



Hopfgartner, Frank (2010) *Personalised video retrieval: application of implicit feedback and semantic user profiles*. PhD thesis.

<http://theses.gla.ac.uk/2132/>

Copyright and moral rights for this thesis are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given



University
of Glasgow

**Personalised Video Retrieval:
Application of Implicit Feedback and
Semantic User Profiles**

Frank Hopfgartner

A thesis submitted in fulfilment of the
requirements for the degree of
Doctor of Philosophy

Department of Computing Science
Faculty of Information and Mathematical Sciences
University of Glasgow

July 2010

Abstract

A challenging problem in the user profiling domain is to create profiles of users of retrieval systems. This problem even exacerbates in the multimedia domain. Due to the Semantic Gap, the difference between low-level data representation of videos and the higher concepts users associate with videos, it is not trivial to understand the content of multimedia documents and to find other documents that the users might be interested in. A promising approach to ease this problem is to set multimedia documents into their semantic contexts. The semantic context can lead to a better understanding of the personal interests. Knowing the context of a video is useful for recommending users videos that match their information need. By exploiting these contexts, videos can also be linked to other, contextually related videos. From a user profiling point of view, these links can be of high value to recommend semantically related videos, hence creating a semantic-based user profile. This thesis introduces a semantic user profiling approach for news video retrieval, which exploits a generic ontology to put news stories into its context.

Major challenges which inhibit the creation of such semantic user profiles are the identification of user's long-term interests and the adaptation of retrieval results based on these personal interests. Most personalisation services rely on users explicitly specifying preferences, a common approach in the text retrieval domain. By giving explicit feedback, users are forced to update their need, which can be problematic when their information need is vague. Furthermore, users tend not to provide enough feedback on which to base an adaptive retrieval algorithm. Deviating from the method of explicitly asking the user to rate the relevance of retrieval results, the use of implicit feedback techniques helps by learning user interests unobtrusively. The main advantage is that users are relieved from providing feedback. A disadvantage is that information gathered using implicit techniques is less accurate than information based on explicit feedback.

In this thesis, we focus on three main research questions. First of all, we study whether implicit relevance feedback, which is provided while interacting with a video retrieval system, can be employed to bridge the Semantic Gap. We therefore first identify implicit indicators of relevance by analysing representative video retrieval interfaces. Studying whether these indicators can be exploited as implicit feedback within short retrieval sessions, we recommend video documents based on implicit actions performed by a community of users. Secondly, implicit relevance feedback is studied as

potential source to build user profiles and hence to identify users' long-term interests in specific topics. This includes studying the identification of different aspects of interests and storing these interests in dynamic user profiles. Finally, we study how this feedback can be exploited to adapt retrieval results or to recommend related videos that match the users' interests. We analyse our research questions by performing both simulation-based and user-centred evaluation studies. The results suggest that implicit relevance feedback can be employed in the video domain and that semantic-based user profiles have the potential to improve video exploration.

– *Leider lässt sich eine
wahrhafte Dankbarkeit mit
Worten nicht ausdrücken.*

Johann Wolfgang von Goethe,
1749-1832

Acknowledgements

For me, studying towards a PhD degree has been an almost four years long journey with many up and downs. I am most grateful to all those people who accompanied me during this time. To most of them, I owe a special “thank you”:

First of all, I thank my supervisory team, Joemon M. Jose, Keith van Rijsbergen and Alan F. Smeaton (Dublin City University) for helpful discussions, guidance and support throughout the past four years. Thank you also to my examiners, Mounia Lalmas and Stefan Rüger (Open University), for positive feedback and an enjoyable viva experience.

Further, I thank past and current members of the Information Retrieval Group for providing an excellent and friendly environment to do research. Special thanks go to Jana Urban (now at Google), Robert Villa, Martin Halvey, David Vallet (now at Autonomous University of Madrid), Erik Graf, Vassilios Stathopoulos, and David Hannah. If you have the chance to collaborate with them, go for it!

A very big “thank you” goes to my “pre-submission reviewers”, Leif Azzopardi, Richard Glassey, Hideo Joho (University of Tsukuba), Iraklis A. Klampanos, Christina A. Lioma (University of Stuttgart), Tamara Polajnar, and Desmond Elliott (now at University of Edinburgh). Thanks guys for helping me bringing this thesis into shape.

I am grateful to my financial supporters, including the European Commission (via their Sixth Framework Programme) and Ebroul Izquierdo. Further, I thank the Multimedia Knowledge Management (MMKM) network, the Government of Andorra and the DELOS Association for various travel grants. Thanks to them, I had the great opportunity to get out of Glasgow from time to time and to meet interesting people who share similar research interests.

Many thanks also go to my friends who accompanied me during the last four years, making my time in Glasgow most enjoyable.

Finally, I dedicate this thesis to my family, especially my parents and my siblings as well as to my grandfather and grandmother (†April 2010) for their unconditional support, tireless love and motivation throughout my studies. Everything I have accomplished, I owe to them.

Table of Contents

1	Introduction	1
1.1	Motivation	2
1.2	Thesis Statement	3
1.3	Contributions	4
1.4	Thesis Outline	5
1.5	Publications	5
2	Background and Related Work	9
2.1	Basic Concepts of Video Retrieval	9
2.1.1	Video Retrieval Overview	9
2.1.2	Structure of Video Documents	10
2.1.3	Video Segmentation	14
2.1.4	Document Representation	16
2.1.5	Interface Designs	22
2.1.6	Summary	33
2.2	Personalised Video Search and Recommendation	33
2.2.1	Personalisation Overview	34
2.2.2	Gathering and Representing Interest	34
2.2.3	Personalisation Techniques	41
2.2.4	State-of-the-art Personalisation Approaches	43
2.2.5	Summary	47
2.3	Evaluation Methodologies	47
2.3.1	Evaluation Overview	47
2.3.2	System-centred Evaluation	48
2.3.3	User-centred Evaluation	52
2.3.4	Simulation-Based Evaluation	57
2.3.5	Summary	62
3	The Role of Implicit Relevance Feedback	63
3.1	Introduction	63
3.2	Low-Level Feedback Events of Video Retrieval Interfaces	65
3.3	User Action Sequences	67
3.3.1	User Action Sequence S_1 (Preview-Click'n'View)	67

3.3.2	User Action Sequence S_2 (Click'n'Browse)	68
3.3.3	User Action Sequence S_3 (Click-View-Explore'n'Slide)	68
3.3.4	User Action Sequence S_4 (Preview-Click-View'n'Browse)	69
3.3.5	User Action Sequence S_5 (Click'n'More)	70
3.3.6	Discussion	71
3.4	Simulation-based Evaluation	72
3.4.1	User Interaction Simulation	72
3.4.2	Parameter Fine Tuning	76
3.4.3	Discussion	79
3.5	Results	80
3.6	Summary	82
4	Exploiting Community-Based Relevance Feedback	84
4.1	Introduction	84
4.2	A Graph-Based Approach for Capturing Implicit Relevance Feedback	86
4.2.1	Labelled Directed Multigraph	87
4.2.2	Weighted Directed Graph	88
4.2.3	Implicit Relevance Pool Recommendation Techniques	90
4.3	System Description	92
4.4	Experimental Methodology	94
4.4.1	Collection and Tasks	94
4.4.2	Experimental Design	96
4.4.3	Participants	96
4.5	Results	97
4.5.1	Task Performance	97
4.5.2	User Exploration	100
4.5.3	User Perception	104
4.6	Summary	109
5	Capturing Long-Term User Information Needs	112
5.1	Introduction	112
5.2	The Need for User Profiles	114
5.2.1	Long-Term Personalisation Scenario	114
5.2.2	A Generic Framework for User Profiling	116
5.2.3	Discussion	118
5.3	Requirements for a User Profile	119
5.3.1	Capturing and Segmenting News Broadcasts	120
5.3.2	Exploiting External Knowledge	121
5.3.3	Categorising News Stories	123

5.3.4	Discussion	124
5.4	Tackling User Profiling Problems	125
5.4.1	User Profile Model	126
5.4.2	Capturing Evolving Interest	128
5.4.3	Capturing Different Aspects of Interest	131
5.4.4	Discussion	132
5.5	Summary	132
6	Simulation-based Evaluation of Long-Term News Video Recommendation Techniques	134
6.1	Introduction	134
6.2	Generating a Daily News Corpus	136
6.2.1	Capturing Daily News	137
6.2.2	Categorising News Stories	138
6.2.3	Semantic Annotation	138
6.2.4	Discussion	140
6.3	Recommendation Techniques	141
6.3.1	Exploiting the Semantic Link	142
6.3.2	Text-Based Recommendation	144
6.3.3	Discussion	145
6.4	Generating Relevance Assessment Lists	145
6.4.1	Assessment Group	146
6.4.2	Gathering of User Interests	148
6.4.3	News Video Assessment	149
6.4.4	Discussion	152
6.5	Generating Simulated User Profiles	153
6.5.1	Training a User Interaction Model	154
6.5.2	Determining Usage Patterns	156
6.5.3	Creating User Profiles	159
6.5.4	Discussion	160
6.6	Simulation-based Evaluation	160
6.6.1	Evaluation Parameters	161
6.6.2	Results	162
6.6.3	Discussion	169
6.7	Summary	169
7	Evaluating Semantic User Modelling	171
7.1	Introduction	171
7.2	System Description	172

7.2.1	Video Processing Component	174
7.2.2	User Interface Component	176
7.2.3	Profiling Component	177
7.2.4	Recommendation Component	178
7.2.5	Discussion	179
7.3	Experimental Methodology	180
7.3.1	Experimental Design	180
7.3.2	Data Corpus	182
7.3.3	Participants	183
7.3.4	Discussion	185
7.4	Results	185
7.4.1	System Usage and Usability	186
7.4.2	Exploiting Implicit Relevance Feedback	188
7.4.3	Recommendation Quality	190
7.4.4	Discussion	192
7.5	Summary	193
8	Conclusion and Future Work	194
8.1	Summary	194
8.2	Conclusion	195
8.2.1	Implicit Indicators of Relevance	195
8.2.2	Collaborative Recommendations	196
8.2.3	Implicit User Profiling	197
8.2.4	Long-Term Recommendation	197
8.3	Limitations	198
8.4	Future Work	199
8.4.1	Implicit Indicators of Relevance	199
8.4.2	Collaborative Recommendations	201
8.4.3	Implicit User Profiling	202
8.4.4	Long-Term Recommendation	204
A	Exploiting Implicit Relevance Feedback: Experimental Documents	237
A.1	Information Sheet	238
A.2	Consent Form	239
A.3	Entry Questionnaire	240
A.4	Post-Search Questionnaire	243
A.5	Exit Questionnaire	247

B	Generating Personalised Ground Truth Data: Experimental Documents	249
B.1	Information Sheet	250
B.2	Technical Information Sheet	251
B.3	Consent Form	252
B.4	Entry Questionnaire	253
B.5	Post-Task Questionnaire	255
B.6	Exit Questionnaire	256
C	Semantic User Modelling for Personalised News Video Access: Experimental Documents	257
C.1	Information Sheet	258
C.2	Simulated Work Task Situation	259
C.3	Consent Form	260
C.4	Entry Questionnaire	261
C.5	Interim Questionnaire	263
C.6	Exit Questionnaire	265

List of Figures

2.1	Video retrieval system framework [Snoek et al., 2007]	16
2.2	Open Video Graphical User Interface (screenshot taken from online system)	24
2.3	Video browsing/retrieval as proposed by Heesch et al. [2004]	25
2.4	Video browsing/retrieval as proposed by Ghoshal et al. [2006]	26
2.5	The Content-based Query Tool as proposed by Rautiainen et al. [2005] .	27
2.6	The Cluster-based Query Tool as proposed by Rautiainen et al. [2005] .	27
2.7	(a) IBM MARVel used for interactive search, and (b) search results grouped by visual clusters [Campbell et al., 2006]	28
2.8	The <i>MediaMagic</i> video search interface Adcock et al. [2008]	29
2.9	The <i>ANSES</i> video search interface [Pickering et al., 2003]	30
2.10	The <i>NewsFlash</i> video search interface [Haggerty et al., 2004]	30
2.11	The <i>Físchlár-News</i> video search interface [Lee et al., 2006]	31
2.12	The <i>NewsRoom</i> video search interface [Diriye et al., 2010]	32
2.13	The <i>Smart Content Factory</i> video search interface [Bürger et al., 2005] .	32
3.1	Possible event combinations on a given document in Sequence S_1 . . .	67
3.2	Possible event combinations on a given document in Sequence S_2 . . .	68
3.3	Possible event combinations on a given document in Sequence S_3 . . .	69
3.4	Possible event combinations on a given document in Sequence S_4 . . .	70
3.5	Possible event combinations on a given document in Sequence S_5 . . .	71
3.6	Steps and related components for a user interaction simulation	74
3.7	Mean Average Precision of the number of documents that users interacted with over up to ten iterations	77
3.8	Precision/Recall of runs with x percent relevant results	78
3.9	Number of Terms for Retrieval	79
3.10	Total number of retrieved relevant shots over all queries using Sequences S_1 – S_5	80
3.11	Mean Average Precision using Sequences S_1 – S_5	81
4.1	Correspondence between the LDM (<i>left</i>) and WDG (<i>right</i>) models . . .	90
4.2	Graph illustrating implicit relevance pool	91
4.3	Interface of the video retrieval system	93

4.4	P@N for the baseline and recommendation systems for varying values of N	98
4.5	Mean average precision (MAP) for baseline and recommendation systems for different groups of user	99
4.6	Average time in seconds to find first relevant shot for baseline and recommendation systems	100
4.7	Probability of a document being relevant given a certain level of interaction. The y-axis represents probability that the video is relevant and the x-axis represents assigned interaction value in our graph	102
5.1	High-level architecture of life-long user modelling	117
5.2	Hierarchy of the concept “Scotland” in DBpedia	122
5.3	Linking “Santorini” and “Majorca” using DBpedia	124
5.4	Effect of different ostensive weighting functions over ten iterations	129
6.1	News Capturing and Segmentation Process	137
6.2	News Video Assessment Interface	151
6.3	Number of relevant rated stories per day (User 7)	152
6.4	Number of topics of interest per day (User 7)	153
6.5	Markov Chain of User Actions	155
6.6	MAP vs. number of documents used for clustering for Semantic Recommendation run	162
6.7	MAP vs. number of documents used for clustering for Named Entity Recommendation run	163
6.8	MAP vs. number of documents used for clustering for Nouns & Foreign Names Recommendation run	164
6.9	MAP vs. query length for Semantic Recommendation run	165
6.10	MAP vs. query length for Named Entity Recommendation run	166
6.11	MAP vs. query length for Nouns & Foreign Names Recommendation run	167
6.12	Recommendation performance of User 6 for every evaluated day with respect to MAP	168
6.13	Recommendation performance of User 6 for every evaluated day with respect to P@5	168
6.14	Recommendation performance of User 4 for every evaluated day with respect to MAP	168
6.15	Recommendation performance of User 4 for every evaluated day with respect to P@5	168
7.1	The conceptual design of the news video recommender system	173

7.2	News Video Recommender Interface	176
7.3	The interface helped me to explore the news collection (lower is better)	187
7.4	The interface helped me to explore various topics of interest (lower is better)	187
7.5	Implicit relevance provided by both user groups	188
7.6	Categories were successfully identified and displayed on the left hand side of the interface (lower is better)	189
7.7	The displayed sub categories represent my diverse interests in various topics. (lower is better)	189
7.8	The displayed sub categories represent my diverse interests in various topics. (lower is better)	190
7.9	The displayed results for each sub category were related to each other. (lower is better)	190
7.10	The displayed results for each category contained relevant stories I didn't receive otherwise. (lower is better)	191
7.11	The displayed results for each category matched with the category description. (lower is better)	191
7.12	The average number of clicks on the sub categories	191
7.13	The average number of manually triggered searches	191

List of Tables

2.1	2×2 latin square design	54
3.1	Weighting of Implicit Features	76
4.1	Values for function $f(a)$ used during the experiment	90
4.2	Event type and the number of occurrences during the experiment	101
4.3	Number of graph elements in graph after each group of four users	101
4.4	Five most popular queries for each task	104
4.5	Perceptions of search process for each system (higher = better)	105
4.6	Perceptions of the retrieval task for each system (higher = better)	106
4.7	Perceptions of the system support for each system (higher = better)	108
4.8	User preferences for the two different systems	109
6.1	Number of entities, concepts and categories in the data collection	140
6.2	The assessors' most popular media types used to consume latest news	147
6.3	The assessors' news topics of interest	147
6.4	Summary of the BBC Online News Assessment Task	149
6.5	Summary of the News Video Assessment Task	150
6.6	The participants' news topics of interest	158
6.7	Probability values of possible action types	159
6.8	Wilcoxon rank-sum test for variable number of stories used for cluster- ing (Semantic vs. Named Entity Recommendations)	165
6.9	Wilcoxon rank-sum test for variable number of stories used for cluster- ing (Semantic vs. Nouns & Foreign names Recommendations)	166
6.10	Wilcoxon rank-sum test for variable number of stories used for cluster- ing (Named Entities vs. Nouns & Foreign Names Recommendations)	167
7.1	Experiment schedule	181
7.2	How the participants retrieve multimedia content	184
7.3	The participants' most popular media types used to consume latest news	184
7.4	The participants' news topics of interest	184
7.5	What the system was used for	186
7.6	News categories that the users were interested in during the experiment	186

– Do not go where the path may
lead, go instead where there is
no path and leave a trail.

Ralph Waldo Emerson,
1803–1882



Introduction

Recently, the Guardian, one of Britain’s most popular daily newspapers published an online article¹⁻¹, recognising the fifth anniversary of the video sharing portal YouTube¹⁻². YouTube is at the forefront of a recent development that, in 2006, convinced the renowned Time magazine to dedicate their Person of the Year award to “You”. You represent the millions of people that started to voluntarily generate (user-generated) content, e.g. in Wikipedia, Facebook and, of course, YouTube. More and more people do not only actively consume content, they have also started to create their own content. Thus, we are observing a paradigm change from the rather passive information consumption habit to a more active information search. Tim Berners-Lee, credited for inventing the World Wide Web, is convinced that eventually this development will completely change the way in which we engage information. During a discussion following his keynote speech “The Web at 20” at Digital Revolution, a documentary due for transmission on BBC Two in 2010, he envisioned¹⁻³ that:

“As a consumer, if I have an internet connection I should be able to get at, and pay for if necessary, anything that has ever been broadcast. So I’m looking forward to when the BBC, for example, [offers] a complete random access library so that I can follow a link, so that somebody can tweet about

¹⁻¹<http://www.guardian.co.uk/technology/2010/apr/23/youtube-five-years-on>, last time accessed on: 7 May 2010

¹⁻²<http://www.youtube.com/>, last time accessed on: 7 May 2010

¹⁻³<http://www.bbc.co.uk/blogs/digitalrevolution/2009/07/tim-bernerslee-and-the-web-at.shtml>, last time accessed on: 7 May 2010

some really, really cool thing or some fun show, or some otherwise boring show, and I can follow a link directly to that. Whether it's pay or free, it's per view, and I get it by following a link, one way or another. I won't be searching channels. I think the concept of a channel is going to be history very quickly on the internet. It's not relevant."

This thesis investigates how users can be assisted in exploring such digital video libraries. The main argument is that personalisation techniques can be employed that assist the users in identifying videos they are interested in. Three main issues are studied within this thesis. First of all, we will study whether users' interests can be identified by interpreting their implicit interactions while interacting with interfaces that provide them access to these libraries. The main challenge we address is how users' multifaceted information needs should be approached, i.e. users' interests in different topics and sub topics. Secondly, we evaluate how this feedback can be preserved over time. This includes both representing users' interests and modelling their evolving interests. Finally, we study how the feedback can be exploited to identify videos that match the users' interests.

In Section 1.1, we motivate this work. Section 1.2 outlines the thesis statement. Section 1.3 outlines the contributions of this work. An overview of the thesis structure is given in Section 1.4. A list of publications that form part of this thesis is given in Section 1.5.

1.1 Motivation

With the growing capabilities and the falling prices of current hardware systems, there are ever increasing possibilities to store and manipulate videos in a digital format. Also with ever increasing broadband capabilities it is now possible to view video online as easily as text-based pages were viewed when the Web first appeared. People are now producing their own digital libraries from materials created through digital cameras and camcorders, and use a number of systems to place this material on the Web, as well as store them in their own individual collections. An interesting research problem is to assist users in dealing with such large and swiftly increasing volumes of video, i.e. in helping them to satisfy their information need by finding videos they are interested in.

The first question that needs to be answered in this context is how users' personal information needs can be identified. In order to identify videos that match this information need, it is helpful to understand the content of the video. However, the difference between the low-level data representation of videos and the higher level concepts users

1.2. Thesis Statement

associate with video, commonly known as the Semantic Gap [Smeulders et al., 2000], provide difficulties. Bridging the Semantic Gap is one of the most challenging research issues in multimedia information retrieval today. A promising method to bridge this gap is to employ relevance feedback (RF) techniques. Relevance feedback can be split into two main paradigms: explicit and implicit relevance feedback. Employing explicit RF, users are asked to judge the relevance of videos. By mining implicit user interaction data, one can infer user intentions and thus could be able to retrieve more relevant information. White [2004] has shown that *implicit relevance feedback* can successfully be employed to support text retrieval tasks. In this thesis, we study whether implicit relevance feedback techniques can be ported into the video domain.

The second question that needs to be addressed is how can this relevance feedback be interpreted. A key prerequisite for this study is how users' interests in multiple aspects can be automatically identified. In this thesis, we approach this question from two different points of view. We argue that users' short-term interests within their current search session should be exploited to satisfy their current information need. Moreover, we argue that users' long-term interests, i.e. interests they expressed over multiple sessions should be considered when identifying interesting video documents. The latter requires storing user feedback in personal profiles. We therefore focus on exploiting both short-term and long-term relevance feedback.

The third question that we address in this thesis is how can the provided feedback be exploited to identify videos of interest. We will study two different approaches. Firstly, we study whether relevance feedback provided by a community of users can be employed to identify relevant videos. Secondly, we argue that setting video documents into their semantic context eases the identification of similar videos. We study this approach by employing a generic ontology that links videos in the collection based on identified concepts within these videos. Ontologies are "content specific agreements" on vocabulary usage and sharing of knowledge [Gruber, 1995].

1.2 Thesis Statement

This thesis aims to study three statements. First of all, we claim that implicit relevance feedback, a well studied technique to adapt or recommend documents in the text retrieval domain, can also be employed in the video domain to bridge the Semantic Gap. Further, we claim that this feedback cannot only be used to recommend videos that match users' interests within one search session, but also videos that match their long-term interests. Finally, we claim that a generic ontology can be used to improve the quality of these recommendations.

The main contributions of this thesis are as follows. Firstly, we identify implicit indicators of relevance by studying representative retrieval systems. Further, we study how these indicators can be employed to provide personalised recommendations within one search session. Finally, we extend this study by introducing a user profiling methodology where users' implicit feedback is stored in structured profiles. Aiming to address our third statement, we introduce our approach of setting video documents into their semantic context by employing an ontology. Exploiting this ontology, we introduce a recommendation technique and evaluate it by employing both simulation-based and user-centred evaluations.

The results that are introduced and discussed in this thesis suggest that implicit relevance feedback is a valid technique in the video domain. Further, they indicate that the introduced implicit user profiling approach and semantic recommendation technique can be successfully employed to provide relevant videos.

1.3 Contributions

This thesis provides the following contributions:

- Implicit indicators of relevance are identified that can be employed in the video domain as a source of implicit relevance feedback.
- An approach for aiding users in the difficult task of video search is introduced. We use community-based feedback mined from the interactions of previous users of our video search system to aid users in their search tasks. This feedback is the basis for providing recommendations to users in our video retrieval system. The ultimate goal of this system is to improve the quality of the results that users find, and in doing so, help users to explore a large and difficult information space and help them consider search options that they may not have considered otherwise.
- A unified architecture for the creation of user profiles is introduced, including feature extraction and representation, reasoning, recommendation and presentation. Discussing various issues arising in the context of long-term profiling, conditions for an implicit user profile capturing users' interests in news videos is outlined.
- A semantic-based user modelling technique to capture users' evolving information needs is introduced. The approach exploits implicit user interaction to capture long-term user interests in a profile. The structured interests are used to retrieve and recommend news stories to the users.

- A novel methodology for the simulation-based evaluation of long-term personalisation techniques is suggested. Individual relevance assessment data is used to simulate users interacting with a video retrieval system.

1.4 Thesis Outline

The remainder of this thesis is structured as follows:

- Chapter 2 surveys related work in the fields of video retrieval, personalised search, recommendation and evaluation techniques.
- Chapter 3 analyses potential implicit indicators of relevance in representative video retrieval systems and evaluates their effect on retrieval performance.
- Chapter 4 introduces a short-term recommendation system that employs implicit relevance feedback of a community of users.
- Chapter 5 outlines the need for implicit user profiles. Further, an approach is introduced that captures users' interests over a longer time period in personalised profiles and exploits these profiles.
- Chapter 6 evaluates the user profiling approach by employing a simulation-based evaluation scheme.
- Chapter 7 aims to confirm the outcome of the previous section by employing a user-centred evaluation scheme.
- Chapter 8 summarises the outcome of this thesis, discusses limitations and suggests future directions.

1.5 Publications

The ideas and results that are presented in this thesis are partly included in various research publications. Further, various papers have been published that lead to the introduced thesis. The corresponding papers are listed in the remainder of this section.

Included Publications

Frank Hopfgartner and Joemon M. Jose. Semantic User Profiling Techniques for Personalised Multimedia Recommendation. *ACM/Springer Multimedia Systems Journal*, 16(4):255–274, 2010.

-
- Frank Hopfgartner and Joemon M. Jose. Semantic User Modelling for Personal News Video Retrieval. In *MMM'10: Proceedings of the 16th International Conference on Multimedia Modeling, Chongqing, China*, pages 336–349. Springer Verlag, 1 2010.
- Klaus Schöffmann, Frank Hopfgartner, Oge Marques, Laszlo Böszörményi, and Joemon M. Jose. Video browsing interfaces and applications: a review. *SPIE Reviews*, 1(1):018004–1–018004–35, 2010. doi: 10.1117/6.00000005.
- Desmond Elliott, Frank Hopfgartner, Teerapong Leelanupab, Yashar Moshfeghi, and Joemon M. Jose. An Architecture for Life-long User Modelling. In *LLUM'09: Proceedings of the Lifelong User Modelling Workshop, Trento, Italy*, pages 9–16, 6 2009.
- Frank Hopfgartner and Joemon M. Jose. On User Modelling for Personalised News Video Recommendation. In *UMAP'09: Proceedings of the First and Seventeenth International Conference on User Modeling, Adaptation, and Personalization, Trento, Italy*, pages 403–408. Springer Verlag, 6 2009.
- Frank Hopfgartner and Joemon M. Jose. *Toward an Adaptive Video Retrieval System*, chapter 6, pages 113–135. Advances in Semantic Media Adaptation and Personalization. CRC Press: Boca Raton, Florida, 2 edition, 2 2009. ISBN 978-1420076646.
- Frank Hopfgartner. Studying Interaction Methodologies in Video Retrieval. *Proceedings of the VLDB Endowment*, 1(2):1604–1608, 2008.
- Frank Hopfgartner, David Vallet, Martin Halvey, and Joemon M. Jose. Search Trails using User Feedback to Improve Video Search. In *MM'08: Proceedings of the ACM International Conference on Multimedia, Vancouver, Canada*, pages 339–348. ACM Press, 10 2008.
- Frank Hopfgartner, David Hannah, Nicholas Gildea, and Joemon M. Jose. Capturing Multiple Interests in News Video Retrieval by Incorporating the Ostensive Model. In *PersDB'08: Proceedings of the Second International Workshop on Personalized Access, Profile Management, and Context Awareness in Databases, Auckland, New Zealand*, pages 48–55. VLDB Endowment, 08 2008.
- Frank Hopfgartner, David Vallet, Martin Halvey, and Joemon M. Jose. Collaborative Search Trails for Video Search. In *JCDL'08: Proceedings of the Joint Conference on Digital Libraries - First International Workshop on Collaborative Search, Pittsburgh, Pennsylvania*, 06 2008.

David Vallet, Frank Hopfgartner, Martin Halvey, and Joemon M. Jose. Community based feedback techniques to improve video search. *Signal, Image and Video Processing: Special Issue on Multimedia Semantics, Adaptation & Personalization*, 2 (4):289–306, 2008.

Frank Hopfgartner and Joemon M. Jose. A News Video Retrieval Framework for the Study of Implicit Relevance. In *SMAP '07: Proceedings of the Second International Workshop on Semantic Media Adaptation and Personalization, London, United Kingdom*, pages 233–236. IEEE, 12 2007.

Frank Hopfgartner and Joemon M. Jose. Evaluating the Implicit Feedback Models for Adaptive Video Retrieval. In *MIR '07: Proceedings of the 9th ACM SIGMM International Workshop on Multimedia Information Retrieval, Augsburg, Germany*, pages 323–331. ACM Press, 09 2007.

Frank Hopfgartner, Jana Urban, Robert Villa, and Joemon M. Jose. Simulated Testing of an Adaptive Multimedia Information Retrieval System. In *CBMI'07: Proceedings of the Fifth International Workshop on Content-Based Multimedia Indexing, Bordeaux, France*, pages 328–335. IEEE, 06 2007.

Jana Urban, Xavier Hilaire, Frank Hopfgartner, Robert Villa, Joemon M. Jose, Siripinyo Chantamunee, and Yoshihiko Gotoh. Glasgow University at TRECVID 2006. In *TRECVID'06: Notebook Papers and Results*, pages 363–367. NIST, 11 2006.

Related Publications

Frank Hopfgartner, Reede Ren, Thierry Urruty, and Joemon M. Jose. *Information Organisation Issues in Multimedia Retrieval using Low-Level Features*, chapter 15. *Multimedia Semantics: Metadata, Analysis and Interaction*. Wiley, 1 edition, 2010. to appear.

Frank Hopfgartner, Thierry Urruty, Pablo Bermejo, Robert Villa, and Joemon M. Jose. Simulated Evaluation of Faceted Browsing based on Feature Selection. *Multimedia Tools and Applications*, 47(3):631–662, 2010.

Hemant Misra, Frank Hopfgartner, Anuj Goyal, P. Punitha, and Joemon M. Jose. News Video Story Segmentation based on Semantic Coherence and Content Similarity. In *MMM'10: Proceedings of the 16th International Conference on Multimedia Modeling, Chongqing, China*, pages 347–357. Springer Verlag, 1 2010.

-
- Teerapong Leelanupab, Frank Hopfgartner, and Joemon M. Jose. User Centered Evaluation of a Recommendation based Image Browsing System. In *IICAI'09: Proceedings of the Fourth Indian International Conference on Artificial Intelligence, Tumkur, India*, pages 558–573. 12 2009.
- Thierry Urruty, Frank Hopfgartner, David Hannah, Desmond Elliott, and Joemon M. Jose. Supporting Aspect-Based Video Browsing – Analysis of a User Study. In *CIVR'09: Proceedings of the ACM International Conference on Image and Video Retrieval, Santorini, Greece*, pages 74–81. ACM Press, 7 2009.
- Frank Hopfgartner, Thierry Urruty, David Hannah, Desmond Elliott, and Joemon M. Jose. Aspect-based Video Browsing – A User Study. In *ICME'09: Proceedings of the IEEE International Conference on Multimedia and Expo, New York, USA*, pages 946–949. IEEE, 6 2009.
- Anuj Goyal, P. Punitha, Frank Hopfgartner, and Joemon M. Jose. Split and Merge based Story Segmentation in News Videos. In *ECIR'09: Proceedings of the 31st European Conference on Information Retrieval, Toulouse, France*, pages 766–770. Springer Verlag, 4 2009.
- Frank Hopfgartner, Thierry Urruty, Robert Villa, and Joemon M. Jose. Facet-based Browsing in Video Retrieval: A Simulation-based Evaluation. In *MMM'09: Proceedings of the 15th International Conference on Multimedia Modeling, Sophia Antipolis, France*, pages 472–483. Springer Verlag, 1 2009.
- Thierry Urruty, Frank Hopfgartner, Robert Villa, Nicholas Gildea, and Joemon M. Jose. A Cluster-based Simulation of Facet-based Search. In *JCDL'08: Proceedings of the Joint Conference on Digital Libraries, Pittsburgh, Pennsylvania*, page 472. ACM Press, 06 2008.
- David Vallet, Frank Hopfgartner, Martin Halvey, and Joemon M. Jose. Community based feedback techniques to improve video search. *Signal, Image and Video Processing: Special Issue on Multimedia Semantics, Adaptation & Personalization*, 2 (4):289–306, 2008.
- David Vallet, Frank Hopfgartner, and Joemon M. Jose. Use of Implicit Graph for Recommending Relevant Videos: A Simulated Evaluation. In *ECIR'08: Proceedings of the 30th European Conference on Information Retrieval, Glasgow, United Kingdom*, pages 199–210. Springer Verlag, 03 2008.

– I find that a great part of the information I have was acquired by looking up something and finding something else on the way.

Franklin P. Adams, 1881–1960

2

Background and Related Work

In this chapter, we provide a survey on related work in the fields of video retrieval, personalisation services and evaluation methodologies. Basic concepts of video retrieval are introduced in Section 2.1. Section 2.2 surveys personalised video search and recommendation techniques. In Section 2.3, we discuss various methodologies that have been established to evaluate research approaches in the information retrieval domain.

2.1 Basic Concepts of Video Retrieval

This section surveys basic concepts of video retrieval. Its foundations are introduced in Section 2.1.1. Section 2.1.2 introduces the classical structure of video documents which highlights the challenges in this domain. Section 2.1.3 introduces approaches to segment video documents for a more efficient document handling. A brief survey on video document representation and on ranking methods is given in Section 2.1.4. Section 2.1.5 introduces representative video retrieval interfaces. Section 2.1.6 summarises and concludes this section. Unless stated otherwise, the material in this section is based on [Blanken et al., 2007; van Rijsbergen, 1979].

2.1.1 Video Retrieval Overview

In recent years, multimedia content available to users has increased exponentially. This phenomenon has come along with (and to much an extent is the consequence of) a rapid development of tools, devices, and social services which facilitate the creation,

storage and sharing of personal multimedia content. A new landscape for business and innovation opportunities in multimedia content and technologies has naturally emerged from this evolution, at the same time that new problems and challenges arise. In particular, the hype around social services dealing with visual content, such as YouTube²⁻¹ or Dailymotion²⁻², has led to a rather uncoordinated publishing of video data by users worldwide [Cunningham and Nichols, 2008]. Due to the sheer amount of large data collections, there is a growing need to develop new methods that support the users in searching and finding videos they are interested in.

Video retrieval is a specialisation of information retrieval (IR), a research domain that focuses on the effective storage and access of data. In a classical information retrieval scenario, a user aims to satisfy their *information need* by formulating a *search query*. This action triggers a retrieval process which results in a list of ranked documents, usually presented in decreasing order of relevance. The activity of performing a search is called the *information seeking* process. A *document* can be any type of data accessible by a retrieval system. In the text retrieval domain, documents can be textual documents such as emails or websites. Image documents can be photos, graphics or other types of visual illustrations. Video documents are any type of moving images. In Section 2.1.2, we introduce the structure of a typical video document. A repository of documents that is managed by an IR system is referred to as a *document collection*. The aim of an IR system is to return relevant documents from the collection with respect to the user's information need. Within this thesis, we will focus our research on news video retrieval, since this is very content reach video material. In Section 2.1.3, we argue that news video documents should be analysed and processed first in order to ease this retrieval process. Preferably, retrieved documents are ranked in accordance to their relevance to the user's information need. In Section 2.1.4, we discuss this retrieval and ranking process further. Aiming to visualise retrieval results for the user, graphical user interfaces are required that allow the user to input their information need and to inspect the retrieved results, thus to access the document collection. Section 2.1.5 surveys state-of-the-art graphical user interfaces in the video retrieval domain. Section 2.1.6 summarises this section.

2.1.2 Structure of Video Documents

Computers serving multimedia and other devices are going to change the handling of videos completely. Films are consistently broadcast, recorded and stored in *digital*

²⁻¹<http://www.youtube.com/>, last time accessed on: 7 May 2010

²⁻²<http://www.dailymotion.com/>, last time accessed on: 7 May 2010

form. In this section, we introduce the typical format of digital video files.

Video Document Format

A video is a sequence of still images, accompanied by an audio stream. Classical digital video standards are the MPEG-1 and MPEG-2 formats. They were released by the Motion Pictures Expert Group (MPEG), the driving force in the development of compressed digital video formats. MPEG-1 videos are often compared to old fashioned VCR recordings. The newer MPEG-2 video format is used to encode videos in DVD quality. It is the standard used for digital television (DVB-T, DVB-S, DVB-C) [Watkinson, 2001].

Another ISO standard that has been defined by MPEG is MPEG-7. Its purpose is to provide a unified standard for the description of multimedia data using meta information. Within this standard, various descriptors have been defined to describe visual content, including colour descriptor, shape descriptor, motion descriptor, face descriptor and textual descriptor [Manjunath et al., 2002]. An overview of MPEG-7 descriptors is given by Manjunath et al. [2002].

Metadata of Multimedia Objects

Besides the video's own text and audio-visual data streams, video documents can be enriched with additional data, the so-called metadata. Blanken et al. [2007] survey various types of metadata that will be outlined in the remainder of this section: (1) A description of the video document, (2) textual annotation and (3) semantic annotations.

Descriptive Data Descriptive data provides valuable information about the video document. Examples are the creation date, director or editor, length of the video and so on. A standard format for descriptive data is called Dublin Core [Weibel, 2005]. It is a list of data elements designed to describe resources of any kind. Descriptive metadata can be very useful when documents within the video collection shall be filtered based on certain document facets. Think, for instance, of a user who wants to retrieve all video documents that have been created within the last month, or all videos from one specific director.

Text Annotations Text annotations are textual descriptions of the content of video documents. More recent state-of-the-art online systems, such as YouTube and Dailymotion, rely on using annotations provided by users to provide descriptions of videos. However, quite often users can have very different perceptions about the same video and

annotate that video differently. This can result in synonymy, polysemy and homonymy, which makes it difficult for other users to retrieve the same video. It has also been found that users are reluctant to provide an abundance of annotations unless there is some benefit to the user [Halvey and Keane, 2007]. van Zwol et al. [2008] approach this problem by transferring video annotation into an online gaming scenario.

Apart from manually created content, an important source for textual annotations are speech transcripts. Huang [2003] argues that speech contains most of the semantic information that can be extracted from audio features. Further, Chang et al. [2005] argue that text from speech data plays an important role in video analysis. In literature, the most common text sources are teletext (also called closed-caption), automatic speech recognition (ASR) transcripts and optical character recognition (OCR) output which will be described in the following. In 1974, the BBC introduced Ceefax, the first teletext system in the world. Its initial purpose was to provide televisual subtitles for the hard of hearing [BBC, 2004]. Nowadays, teletext is used by many broadcasting stations as an additional service. Teletext provides a good transcript of the broadcast. Another popular text source is the output of ASR tools. Though such tools are still far from being perfect, Chang et al. [2005] argue that ASR has been useful in video retrieval context. A third text source is the recognition of visual text in images and videos. According to Lienhart [2003], texture-based object recognition can be divided into two categories:

- Scene text is text which appears as part of a scene, i.e. street names or text on T-Shirts. This type of text is hard to detect and has rarely been studied.
- Overlay text is text which has been included into the scene. It often contains important information and is therefore suitable for retrieval purposes.

Considering that textual annotations can be a valuable source for IR systems aiming to retrieve the video documents, various approaches have been studied to automatically determine textual annotations. Note that due to the Semantic Gap (see Section 2.1.4), automatically annotating video images is a non-trivial problem. A survey of state-of-the-art approaches is given by Magalhães and Rüger [2006]. More recent examples include [Stathopoulos and Jose, 2009; Llorente and Rüger, 2009; Llorente et al., 2008; Qi et al., 2007; Wang et al., 2007b].

Semantic Annotations Another type of annotations are semantic annotations. The idea is here to identify concepts and define their relationship between each other and the video document. Concepts can hence set the content of video documents into a semantic context. This is especially useful for semantic retrieval approaches. We will survey such approaches in Section 2.1.4. The previously mentioned MPEG-7 standard

allows for describing multimedia documents and their semantic descriptions. Promising extensions include COMM (Core Ontology for Multimedia), an ontology introduced by [Arndt et al. \[2007\]](#). Ontologies are “content specific agreements” on vocabulary usage and sharing of knowledge [[Gruber, 1995](#)]. Other metadata models include [Durand et al. \[2005\]](#); [Tsinaraki et al. \[2005\]](#); [Bertini et al. \[2007\]](#), who aim to enrich interactive television broadcast data with additional information by combining existing standards. All approaches build hence upon similar ideas.

Semantic annotations can either be derived from textual annotations or from the videos’ low-level features, i.e. by identifying high-level concepts. [Magalhães and Rüger \[2006\]](#) provide a survey on state-of-the-art methodologies to create semantic annotations for multimedia content. They distinguish between three semantic annotation types: (1) hierarchical models, (2) network models and (3) knowledge-based models. *Hierarchical* models aim to identify hierarchical relations or interdependencies between elements in an image or key frame. Examples include [[Barnard and Forsyth, 2001](#); [Blei and Jordan, 2003](#)]. *Network models* aim to infer concepts given the existence of other concepts. Surveyed approaches are [[Kumar and Hebert, 2003](#); [He et al., 2004](#)]. The third approach, *knowledge-based* models relies on prior knowledge to infer the existence of concepts. [Bürger et al. \[2005\]](#), for example, enrich news video data with a thesaurus of geographic names. Therefore, they determine location names within the news reports’ transcripts and map these with their thesaurus. Further, they identify thematic categories by mapping terms in the transcript with a controlled vocabulary. A similar approach is introduced by [Neo et al. \[2006\]](#), who use the WordNet lexical database [[Fellbaum, 1998](#)] to semantically enrich news video transcripts. Even though their approaches allow linking of related news videos, the main problem of their approaches is text ambiguity. Other examples include [[Tansley, 2000](#); [Simou et al., 2005](#)].

Discussion

In this section, we introduced the format of video documents. As we have shown, videos consist of a set of audio-visual signals and accompanying metadata. We argued that the audio-visual features can be described by low-level feature descriptors, the main description standard being MPEG-7. Retrieving videos using these low-level features is, due to the Semantic Gap, a challenging approach. Consequently, other sources are required to support video retrieval. We surveyed three different sources that can be summarised as metadata: (1) Descriptive Data, (2) Text Annotations and (3) Semantic Annotation. All approaches aim to provide annotations in textual form that allow to bridge the Semantic Gap. Within this thesis, we will study various research questions

and methodologies using news video collections. News video are of content rich nature. Therefore, we will base our study on using this rich metadata.

2.1.3 Video Segmentation

As we will show in Section 2.1.4, IR systems index documents and retrieve these documents based on their relevance to a search query. This approach is problematic, however, if a document contains short paragraphs which are highly relevant to the query, while the majority of the document is not relevant at all. Classical ranking algorithms, e.g. based on the document's term frequency, will result in a low ranking of this document. A promising approach to tackle this problem in the text domain is to split documents into shorter *passages* (e.g. [Salton et al. \[1993\]](#)). Various potential advantages arise when considering these passages as unit of retrieval results. First of all, individual passages will be ranked higher than documents which contain the corresponding passages. Consequently, retrieval performance increases. Second, ranking problems due to variable document lengths is minimised, assuming that passages have a similar length. Third, short passages are easier to assess for the user than long documents. Users can easily browse through short results to search for their information need. The same problems apply to videos in the news video domain. However, due to the different nature of news video documents, successful passage retrieval approaches cannot easily be adopted. This section introduces typical segmentation of news videos.

Shot Segmentation

The atomic unit of access to video content is often considered to be the video *shot*. [Monaco \[2009\]](#) defines a shot as a part of the video that results from one continuous recording by a single camera. It hence represents a continuous action in time and space in the video. Especially in the context of professional video editing, this segmentation is very useful. Consider for example a journalist who has to find shots in a video archive that visualise the context of a news event. Shot segmentation infers shot boundary detection, since each shot is delimited by two consecutive shot boundaries. [Hanjalic \[2002\]](#) provide a comprehensive overview on issues and problems involved in automatic shot boundary detection. A more recent survey is given by [Smeaton et al. \[2010\]](#).

News Story Segmentation

A more consumer-oriented approach is to segment videos into semantically coherent sequences. For instance, sports fans want to watch specific highlights of a game rather than short shots depicting only parts of this highlight.

In the news video domain, such coherent sequences are news stories. News stories are commonly seen as segments of a news broadcast with a coherent news focus which contain at least two independent declarative clauses. News bulletins consists of various continuous news stories, such as reports about political meetings, natural disasters or sports events. [Chaisorn and Chua \[2002\]](#) argue that the internal structure of news stories depends on the producer's style. While some stories consist of anchor person shots only, often with a changing background image, other stories can consist of multiple different shots, e.g. other anchor persons, graphics or animations, interview scenes or shots of meetings.

News story segmentation is essentially finding the boundaries where one story ends and the other begins. Various text-based, audiovisual-based and combinations of all features have been studied to segment news videos accordingly. Detailed surveys are given by [Arlandis et al. \[2005\]](#) and [Chua et al. \[2004\]](#).

Discussion

In this section, we introduced typical approaches to segment news videos into smaller units. By considering these segments as unit of retrieval, segmenting videos into smaller chunks is a necessary requirement to ease both retrieving and exploring video document collections. The level of segmentation depends on the actual IR task. The most common segmentation approach is to automatically identify video *shots*, i.e. scenes that have been recorded using one camera only. Professional video editors, for example, might have an information need that can be satisfied by retrieving video shots. Another, more challenging approach is to identify news *stories*. Retrieving news stories can be the aim of a personal IR task, i.e. a user might be interested in certain latest news that are represented in news stories.

Note that under some conditions, e.g. when content providers enrich their video data with time stamps indicating segmentation boundaries, the task of identifying these boundaries becomes trivial. Video segmentation can then be seen as a simple processing task rather than a complex research challenge.

Within this thesis, we will study two different IR scenarios that require both segmentation types. Since video segmentation is not the main focus of this work though, we will not discuss segmentation techniques in detail. For further details, the reader is referred to the respective publications within this section.

2.1.4 Document Representation

Video retrieval systems aim to retrieve relevant video documents that match the users' information need. Various conditions need to be fulfilled to enable this process. Snoek et al. [2007] sketched a common framework that applies for most state-of-the-art video retrieval engines. As presented in Figure 2.1, the framework can be divided into an indexing engine and a retrieval engine.

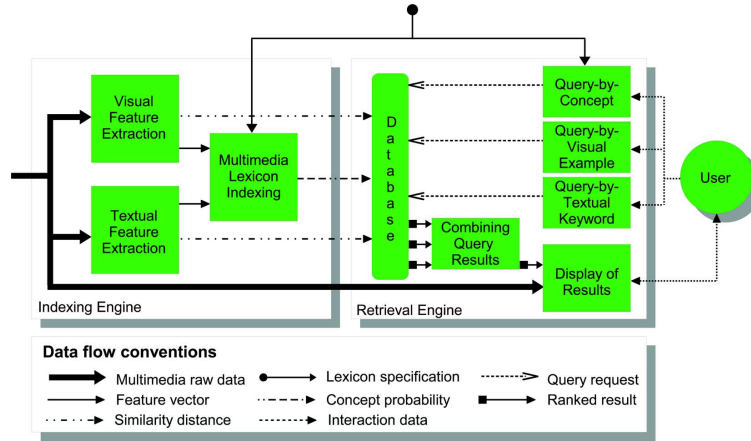


FIGURE 2.1: Video retrieval system framework [Snoek et al., 2007]

The first component involves the indexing of the video data, so that documents can be retrieved that match the users' information need. With respect to current systems, this indexing can be incorporated on a visual, textual and semantic level. Most video retrieval engines store their indexed data collection in a database. Techniques are required to match both information need and the video collection. These tasks are fulfilled by the retrieval engine. In the remainder of this section, we briefly introduce state-of-the-art approaches that address these issues.

Video Indexing

Video indexing is the backbone of all video retrieval engines. Indexing approaches aim to develop effective and efficient methodologies for storing, organising and accessing video contents. As we have shown in Section 2.1.2, a video document consists of several modalities, e.g. a video document is made up of audio tracks, visual streams and different types of annotations. Thus, video indexing has to take numerous modality features into consideration. Moreover, these features are of various nature. Video indexing techniques can be split into three main categories: content-based indexing, text-based indexing and semantic indexing. Note that we will not focus on video indexing within this thesis, since it is out of scope of this work. In Hopfgartner et al. [2010a], we present

two multimedia indexing approaches that aim to address the two main challenges in content-based indexing: (1) the high dimensional feature space of multimedia data and (2) the variable character of feature dimensions, i.e. boolean and multi-value features. A survey on content-based video indexing is given by [Baeza-Yates and Ribeiro-Neto \[1999\]](#), semantic indexing is surveyed by [Snoek and Worring \[2005\]](#).

Retrieval

As discussed before, video data consists of multimodal features, including text, audio visual features and semantic annotations. Consequently, there are a number of different ways in which a user can query a video retrieval system. As [Snoek et al. \[2007\]](#) pointed out, three query formulation paradigms exist in the video retrieval domain: query-by-textual-keyword, query-by-visual-example and query-by-concept.

Query-By-Textual-Keyword Query-by-textual-keyword is one of the most popular methods of searching for video [\[Hauptmann, 2005\]](#). It is simple and users are familiar with this paradigm from text-based searches. Query-by-text relies on the availability of sufficient textual descriptions, including descriptive data, transcripts and annotations. A survey on these data is given in Section [2.1.2](#).

Query-By-Visual-Example Query-by-visual-example has its roots in content-based image retrieval. It allows the users to provide sample images or video clips as examples to retrieve more results. This approach uses the low-level features that are available in images and videos, such as colour, texture and shape to retrieve results. The basic idea is that visual similarity can be used to identify relevant documents. Content-based image retrieval has been well studied, a survey is given by [Aigrain et al. \[1996\]](#).

The main problem of content-based retrieval is that users have difficulties mapping high-level concepts and low-level features of images. One reason for this is the subjectivity of human perception [\[Rui et al., 1998a\]](#). Different persons, or the same person in different situations, may interpret visual content differently. For example, one person may focus on an image's shape feature, while another one focuses on its colours. Even if focusing on the same feature, the perception of similar images can differ. [Smeulders et al. \[2000\]](#) refer to this problem as the Semantic Gap:

“The Semantic Gap is the lack of coincidence between the information that one can extract from the sensory data and the interpretation that the same data has for a user in a given situation.”

Bridging the Semantic Gap is considered one of the most challenging research issues in multimedia information retrieval today [Jaimes et al., 2005]. Differing from image content, videos provide additional features which can be used for content-based retrieval, including motion and audio features. A survey on content-based video retrieval is provided by Gurrin [2009].

Query-By-Concept In an attempt to bridge the Semantic Gap, a great deal of interest in the multimedia search community has been invested in query-by-concept, also referred to as concept-based, conceptual or semantic retrieval. A survey on concept-based retrieval is given by Snoek and Worring [2009]. Conceptual retrieval relies on semantic annotations, i.e. high level concepts which have been associated with the video data. A well known set of high-level semantic concepts has been explored by the Large Scale Ontology for Multimedia (LSCOM) initiative [Naphade et al., 2006], a subset of which is used within TRECVID to study concept-based retrieval. Considering semantic concepts as additional textual annotation, documents can be retrieved by triggering textual search queries. Hildebrand et al. [2007] analysed state-of-the-art semantic retrieval systems, concluding that semantic concepts are often used to filter retrieval results. Query-by-concept is an extension to both query-by-textual-keyword and query-by-visual-example, narrowing down corresponding results. Indeed, the most successful video retrieval systems that have been evaluated within TRECVID (e.g. [Snoek et al., 2008; Hauptmann et al., 2005]) employ these two approaches to improve their retrieval results.

Ranking Methods

After introducing different approaches to express users' interests in a search query, this section surveys how documents within the data collection are matched to the search query. Note that within this work, we will exploit textual annotation of the video data for retrieval purposes. We therefore focus on text-based ranking methods only.

Boolean Models A very basic model to match documents to a query is the *Boolean model* [Salton et al., 1983]. It relies on boolean logic and thus requires an exact matching of document and query terms. Retrieved documents are returned without a relevance rating. Also due to this lack of ranking, Boolean models are no longer considered as a state-of-the-art ranking method [Zobel and Moffat, 2006].

Vector Space Models Salton et al. [1975] argued for the representation of terms as vectors in a multi-dimensional (linear) space. Aiming to match documents to search

queries, both are represented as vectors of terms. Documents are retrieved by computing the distance between both vectors; the closer the distance between the vectors, the more relevant is the document to the search query. Various techniques have been studied to compute the distance between vectors. A survey on these approaches is given by [van Rijsbergen \[1979\]](#).

Probabilistic Models An alternative ranking method estimates the probability of a document being relevant to a given query. It is based on the *probability ranking principle*, which was formalised by [Maron and Kuhns \[1960\]](#). Within this model, documents in a collection are considered to be either relevant or non-relevant to an information interest expressed by a search query. Documents should then be ranked based on their probability of being relevant:

$$\frac{P(\vec{d}|r)}{P(\vec{d}|n)} \quad (2.1)$$

where \vec{d} is a document, r indicates relevance of the document, n indicates non-relevance of the document and $P(\vec{d}|r)$ is the probability of document \vec{d} being relevant. A survey on probabilistic ranking models is given by [Sparck-Jones et al. \[2000\]](#). In the remainder of this section, we introduce one of the most popular probabilistic models, the well-known Okapi BM25 model which was introduced by [Robertson et al. \[1994\]](#).

Within this model, each document \vec{d} in a collection C is defined as a vector $\vec{d} = (d_1, \dots, d_v)$, where d_j represents the *term frequency* of the j^{th} term in \vec{d} , whereas V is the total number of terms in the vocabulary that is used within the collection. BM25 computes a score $w_j(\vec{d}, C)$ that can be simplified as

$$w_j(\vec{d}, C) = \frac{(k_1 + 1)d_j}{k_1((1 - b) + b\frac{dl}{avdl}) + d_j} \log \left(\frac{N - df_i + 0.5}{df_j + 0.5} \right), \quad (2.2)$$

where dl represents the document length and $avdl$ is the averaged document length, df_j is the document frequency of term j , N is the number of documents in the collection, and b and k_1 are parameters. b controls the document length normalisation and k_1 controls the non-linear term frequency effect. The recommended values are $b = 0.75$ and $k_1 = 1.2$. Hence, BM25 computes a matching score of a document \vec{d} for a given search query q as:

$$W(\vec{d}, q, C) = \sum_j w_j(\vec{d}, C) \cdot q_j \quad (2.3)$$

[Zaragoza et al. \[2004\]](#) introduce an extension of this ranking function, referred to as BM25F. Within this model, the structure of documents are considered. It is especially

useful when a higher importance should be given to terms appearing in different fields of a document, e.g. in the title. Within this model, documents can consist of a collection of field types $T = \{1, \dots, f, \dots, K\}$, where $f = 1$ may stand for the document's Title, $f = 2$ may be the document's Abstract, and so on. Considering these fields, a document \mathbf{d} can be defined as a vector of text-fields:

$$\mathbf{d} = (\vec{d}[1], \vec{d}[2], \dots, \vec{d}[f], \dots, \vec{d}[K]). \quad (2.4)$$

$\vec{d}[1]$, for example, might represent the title field of document \mathbf{d} , while $\vec{d}[2]$ might represent the abstract. The collection of these structured documents is defined as \mathbf{C} . Each field $\vec{d}[f]$ is a vector of term frequencies $(d[f]_j)_{j=1..V}$. Aiming to weight each field differently, the field weight \mathbf{v} is defined as a vector $\mathbf{v} \in \Re^K$. Treating each field type as a separate collection, the BM25 weighting function (2.3) can be applied to combine these collections:

$$W(\vec{d}[f], q, C) = \sum_j w_j(\vec{d}[f], C) \cdot q_j \quad (2.5)$$

However, in order to not only consider the content of the collection, but also the field structure of the documents, $W(\vec{d}[f], q, C)$ needs to be extended into a new function $W(\mathbf{d}, q, \mathbf{C}, \mathbf{v})$, e.g. by forming a linear combination:

$$W_1(\mathbf{d}, q, \mathbf{C}, \mathbf{v}) = \sum_{f=1}^K v_f \cdot W(\vec{d}[f], q, C) \quad (2.6)$$

Diversity Ranking Above introduced methods rank retrieval results based on their relevance to a given query. Similar documents, i.e. documents depicting the same features, will appear in proximity to each other in such ranked lists. Another approach is to *diversify* retrieval results. Diversity ranking techniques aim to present the users with a wider range of options in their search results by presenting a diverse set of results that embody many possible interpretations of the users query.

Some initial work has been carried out in the area of diversification in text retrieval. [Zhang et al. \[2005\]](#) diversify search results in the context of Web search. They propose a novel ranking scheme named Affinity Ranking which re-ranks search results by what they call diversity and information richness. The TREC Novelty Track [[Soboroff, 2004](#)] aimed to encourage research in finding novel sentences in a set of relevant sentences.

There have been very few studies of diversity measures for image and video search. [Song et al. \[2006\]](#) acknowledge the need for diversity in search results for image retrieval. They propose a re-ranking method based on topic richness analysis to enrich

topic coverage in retrieval results, while maintaining acceptable retrieval performance. More recently, [van Zwol et al. \[2008\]](#) propose a diversification model for image search results that is based on annotations associated with an image. The contribution of this work is two-fold. Firstly, the diversity is a result of the retrieval model and not a post retrieval step. Secondly, they balance precision and diversity by estimating the query model from the distribution of tags which favours the dominant sense of the query. While this approach is shown to be useful, it suffers from the lack of annotations which is common for multimedia is shared online [[Halvey and Keane, 2007](#)]. Although not on diversification, there has been some work carried out looking at the role of dissimilarity in image retrieval. [Hu et al. \[2008\]](#) look at a number of different methods for calculating dissimilarity. They evaluate the performance for a number of different measures for a number of different feature spaces for the Corel collection. Based on these evaluations, they identify a number of the best dissimilarity measures. [Halvey et al. \[2009c\]](#) introduce a number of methods to promote diversity in video search results by performing a user study. They observed that users perceived diversity ranked result as the most appropriate and complete results.

Discussion

In this section, we introduced three main concepts that are required to retrieve videos. The first step includes indexing the videos documents, hence creating an index of the whole video collection. We argued that, due to the multimodal nature of video documents, indexing is a non-trivial task. Further, we introduced three different querying approaches that allow the user to express their information need: Query-by-textual-keyword, Query-by-visual-example and Query-by-concept. Within this thesis, we will rely on query-by-textual-keyword and query-by-concept. Nevertheless, we consider example-based search queries as an important aspect of video retrieval. Thus, extended work could include this querying paradigm as well. Finally, we surveyed how indexed documents can be matched to the search query. [Yan and Hauptmann \[2007\]](#) analysed the most common approaches of ranking video documents on text-based queries. They summarise that the Okapi BM25 model outperforms vector space models. We therefore rely on this classical ranking method within this work. Further, we introduced methods to diversify retrieval results. Even though a diverse representation of retrieval results allows users to faster explore a document collection, we neglect this ranking method within this thesis, since this would extend the focus of this work. Diversifying results, however, is an interesting research question and can be considered in future work.

Another important concept in interactive video retrieval is the design of graphical

user interfaces that allow the users to both express their information need and to interact with the retrieval results. We survey state-of-the-art video retrieval interfaces in the next section.

2.1.5 Interface Designs

According to [Spink et al. \[1998\]](#), users are often uncertain of their information need and hence have problems finding a starting point for their information seeking task. And even if users know exactly what they are intending to retrieve, formulating a “good” search query can be a challenging task. This problem even exacerbates when dealing with multimedia data. Graphical user interfaces serve here as a mediator between the available data corpus and the user. It is the retrieval systems’ interface which will provide users facilities to formulate search queries and/or to dig into the available data. [Hearst \[2009\]](#) outlines various conditions that dominate the design of state-of-the-art search interfaces. First of all, the process of searching is a means toward satisfying an information need. Interfaces should therefore avoid being intrusive, since this could disturb the users in their seeking process. Moreover, satisfying an information need is already a mentally intensive task. Consequently, the interface should not distract the users, but rather support them in their assessment of the search results. Especially in the WWW domain, search interfaces are not used by high expertise librarians only but also by the general public. Therefore, user interfaces have to be intuitive to use by a diverse group of potential users. Consequently, widely used web search interfaces such as Google, Bing or Yahoo consist of very simple interfaces, mainly consisting of a keyword search box and results being displayed in a vertical list.

Considering the success of above mentioned web search engines, it is not premature to assume that these interfaces effectively handle the interaction between the user and the underlying text retrieval engine. However, text search engines are rather simple in comparison to their counterparts in the video retrieval domain. Therefore, [Jaimes et al. \[2005\]](#) argue that this additional complexity introduces further challenges in the design of video retrieval interfaces.

The first challenge is how users shall be assisted in formulating a search query. As shown in the previous section, three query formulation paradigms dominate the video retrieval domain: Query-by-textual-keyword, Query-by-visual-example and Query-by-concept. Video retrieval interfaces need to be provided with corresponding query formulation possibilities in order to support these paradigms. Another challenge is how videos shall be visualised allowing the user an easy understanding of the content. In the text retrieval domain, short summaries, referred to as snippets, are usually displayed

which allow the users of the system to judge the content of the retrieved document. Multiple research (e.g. [Tombros and Sanderson \[1998\]](#); [White et al. \[2003\]](#)) indicate that such snippets are most informative when they show the search terms in their corresponding context. Considering the different nature of video documents and query options, identifying representative video snippets is a challenging research problem. Moreover, another challenge is how users can be assisted in browsing the retrieved video documents. Systems are required which enable users to interactively explore the content of a video in order to get knowledge about its content.

In [[Schöffmann et al., 2010](#)], we survey representative video browsing and exploration interfaces. We distinguish between interfaces that support video browsing that rely on interaction similar to classical video players, video exploration interfaces and unconventional video visualisation interfaces. Another survey is given by [Snoek and Worring \[2009\]](#), who focus on concept-based video retrieval interfaces. In this section, we survey state-of-the-art video retrieval interfaces with respect to the following features: (1) Interfaces that consider video *shots* as the basic unit of retrieval and (2) interfaces where video *stories* are main unit of retrieval. Note that we will focus on news video retrieval interfaces only, since most interfaces have been developed for news video collections. Besides, our research is focused on news video retrieval as well.

Shot-Based Video Retrieval Interfaces

In one of the earlier efforts for supporting video retrieval, [Arman et al. \[1994\]](#) proposed to use the concept of *key frames* (denoted as *Rframes* in their paper), which are representative frames of shots, for chronological browsing the content of a video sequence. For every shot a key frame is selected based on an analysis of low-level features. In addition to chronological browsing of key frames, their approach already allows selecting a key frame and searching for other similar key frames in the video sequence. Several other papers have been published that use key frame based browsing of shots in a video sequence, usually by showing a page-based grid-like visualisation of key frames (this is also called *Storyboard*) [[Yeung and Yeo, 1997](#); [Zhang et al., 1997](#); [Komlodi and Marchionini, 1998](#); [Komlodi and Slaughter, 1998](#); [Ponceleon et al., 1998](#); [Wilcox et al., 1999](#); [Sull et al., 2001](#); [Geisler et al., 2002](#)]. Some of them propose clustering of key frames into a hierarchical structure [[Zhang et al., 1997](#); [Ponceleon et al., 1998](#); [Sull et al., 2001](#)]. Considering the large number of systems that visualise search results in a storyboard view, this approach can be seen as the standard visualisation method. In the remainder of this section, we survey few representative interfaces which rely on this visualisation paradigm. An introduction on other paradigms is given by [Christel](#)

[2008].

First efforts to provide a digital library started in 1996. The researchers from the University of North Carolina at Chapel Hill indexed short video segments of videos and joined them with images and text and hyperlinks in a dynamic query user interface. Their project evolved since then so that now, digitalised video clips from multiple sources are combined to the Open Video Project [Geisler et al., 2002]. Figure 2.2 shows a screenshot of the actual interface. It allows to trigger a textual search by entering a query, denoted (1) in the screenshot and the possibility to browse through the collections. Results are listed based on their importance to the given search query, denoted (2) in the screenshot.

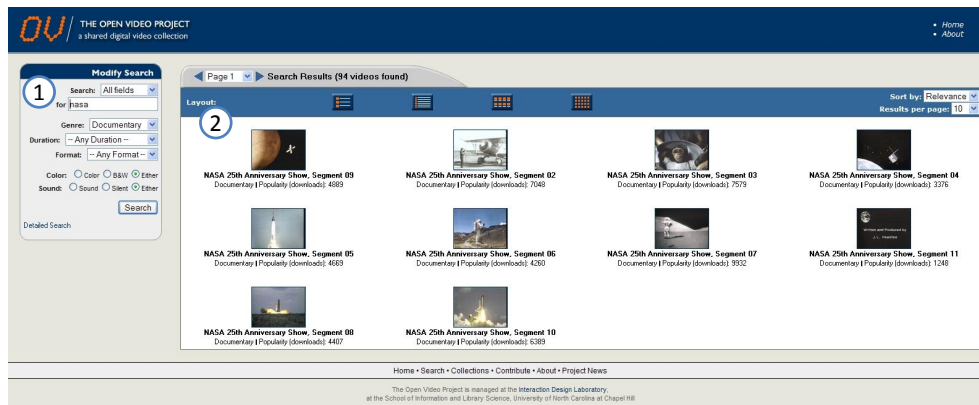


FIGURE 2.2: Open Video Graphical User Interface (screenshot taken from online system)

With the *CueVideo* project, Srinivasan et al. [1999] have presented a browsing interface which allows several visualisations of the video content. Their system is based on shots and consists of visual content presentation, aural content presentation, and technical statistics. Visual content presentation comprises (1) a storyboard where for each shot a key frame is presented, and a (2) *motion storyboard* where for each shot an animated image is presented. The *audio view* shows a classification of the audio tracks into the categories music, speech, and interesting audio events. In a user study they found out that the most popular view was the storyboard view, which is a similar result as already found by Komlodi and Marchionini [1998].

Heesch et al. [2004] presented a tool for video retrieval and video browsing (Figure 2.3). The tool allows searching and browsing a video in different dimensions in a storyboard manner. A user can (1) select an image (or key frame of a shot) as input. This image is further used by a feature-based search (2) that uses a feature-vector consisting of nine different features for comparison (in general colour, texture, and transcript text). A user can manually tune the weighting of the different features. In the right part of the window, the results of the search are presented in a line-by-line and page-by-page

2.1. Basic Concepts of Video Retrieval

manner (3) . The best result is presented at the top-left position of the first page and the worst result is presented at the bottom-right position of the last page. Furthermore, they use a relevance feedback technique in order to improve repeated search. On another tab (called NN^k network, (4)), the nearest neighbours of a selected key frame can be shown in a graph-like visualisation. To provide temporal browsing they also use a *fish-eye visualisation* at the bottom of the window (5) in which the *image-of-interest* (selected on any view) is always shown in the center.

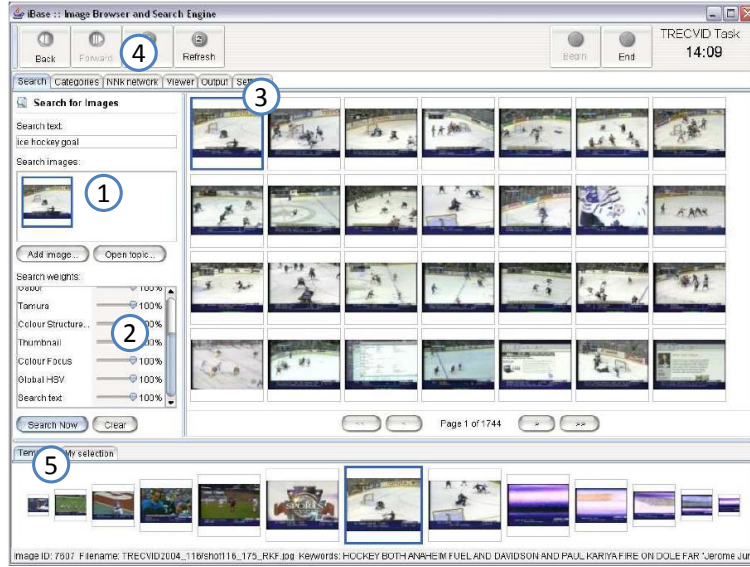


FIGURE 2.3: Video browsing/retrieval as proposed by Heesch et al. [2004]

An extension of this approach is introduced by Ghoshal et al. [2006]. Their interface, shown in Figure 2.4, is split into two main panels with the browsing panel taking up to 80% of the screen. The browsing tab (1) is divided into four tabs that provide different categories: Image & Feature Search, Content Viewer, Search Basket and NN^k key frame browsing. In the Image & Feature Search tab (2), users can enter free text, named entities and visual concepts. Besides, they can specify the weighting of each textual and visual feature in using a sliding bar (3). The Content Viewer tab is divided into two tabs. On the left hand side (4), textual metadata of the last clicked key frame is presented while on the right hand side, the full key frame is shown. In the Search Basket tab, key frames which are currently selected are displayed. The NN^k browsing tab shows these thirty key frames which are nearest to the last clicked key frame in the visual feature space.

Rautiainen et al. [2005] study content-based querying enriched with relevance feedback by introducing a content-based query tool. Their retrieval system supports three different querying facilities: query-by-textual-keyword, query-by-example and query-

2.1. Basic Concepts of Video Retrieval



FIGURE 2.4: Video browsing/retrieval as proposed by Ghoshal et al. [2006]

by-concept. The interface, shown in Figure 2.5, provides a list of semantic concepts a user can choose from. Textual-based queries can be added in a text field on the top left hand side of the interface. Retrieved shots are represented as thumbnails of key frames, together with the spoken text in the most dominant part of the interface. By selecting key frames, users can browse the data collection using a cluster-based browsing interface [Rautiainen et al., 2004]. Figure 2.6 shows a screenshot of this interface. It is divided into two basic parts: On top is a panel displaying the selected thumbnail and other frames of the video in chronological order (1). The second part displays similar key frames which have been retrieved by multiple content-based queries based on user-selected features (2). The key frames are organised in parallel order as a similarity matrix, showing the most similar matches in the first column. This enables the user to browse through a timeline and see similar shots at the same time. Each transition in the timeline will automatically update the key frames in the similarity matrix.

Campbell et al. [2006] introduce a web-based retrieval interface. Using this interface, users can start a retrieval based on visual features, textual queries or concepts. Figure 2.7a shows an example retrieval result. The interface provides functionalities to improve the visualisation of retrieved key frames by grouping them into clusters according to their metadata, such as video name or channel. Figure 2.7b shows an example grouping.

A similar approach is studied by Bailer et al. [2006]. In their interface, retrieval results are categorised into clusters. Single key frames represent each cluster in the result list (1). Controls around the panel (2) depicting the search results allow the users

2.1. Basic Concepts of Video Retrieval

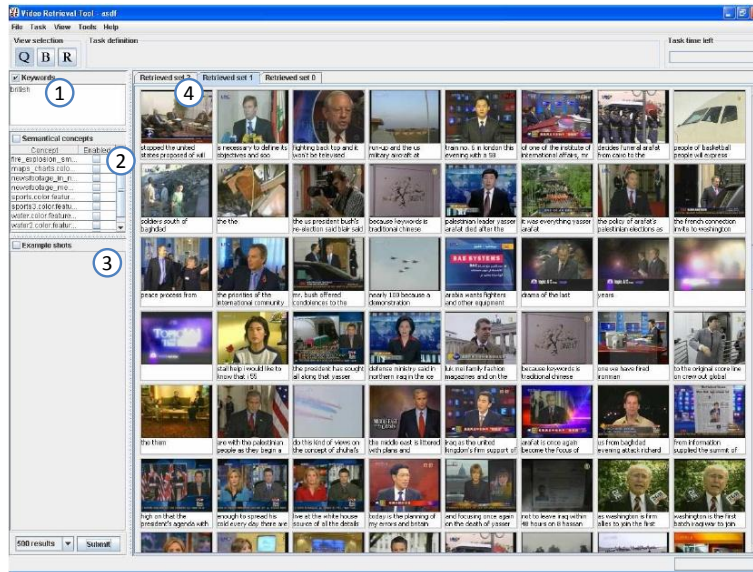


FIGURE 2.5: The Content-based Query Tool as proposed by Rautiainen et al. [2005]

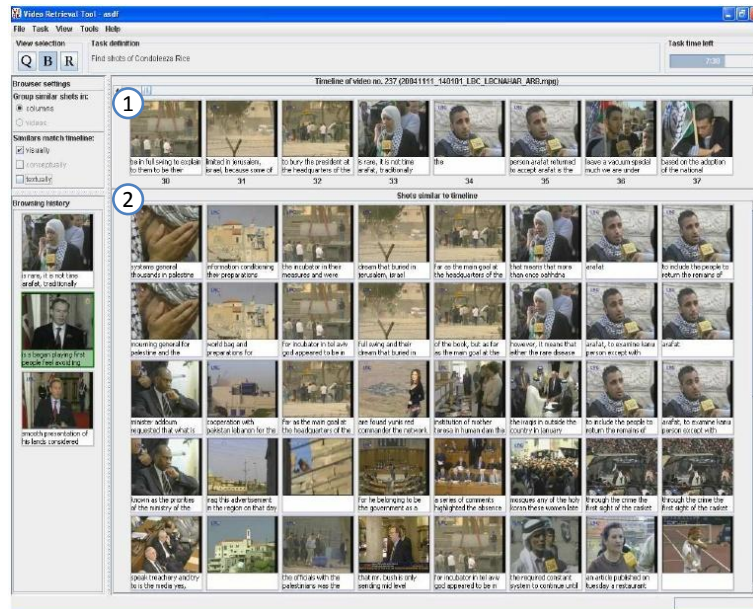


FIGURE 2.6: The Cluster-based Query Tool as proposed by Rautiainen et al. [2005]

to resize the presentation of these key frames and to scroll through the list.

Villa et al. [2008b] presented the *FacetBrowser*, a Web-based tool which allows performing simultaneous search tasks within a video. A similar approach is evaluated by Hopfgartner et al. [2009]. The idea behind is to enable a user to explore the content of a video by individual and parallel (sub-)queries (and associated search results) in a way of exploratory search. A facet in that context is modeled as an individual search among others. The tool extracts speech transcripts from shots of the video for textual

2.1. Basic Concepts of Video Retrieval



FIGURE 2.7: (a) IBM MARVel used for interactive search, and (b) search results grouped by visual clusters [Campbell et al., 2006]

search. The results of a query are shown in a storyboard view where, in addition, a list of user-selected relevant shots for a particular query is shown as well. Moreover, the interface allows to add/remove search panels, to spatially move search panels, and to reuse search queries already performed in the history of a session.

Adcock et al. [2008] presented an interactive video search system called *Media-Magic*. The shot-based system allows searching at textual, visual, and semantic level. They use shot detection and colour correlograms to analyse the content. A rich search interface is provided, which enables searching text queries, image queries and concept queries. As given in Figure 2.8, (3) shows the search topic and supporting example images. At (2) a user can enter textual or image queries. The results are presented in a storyboard visualisation in area (1). (5) shows a "selected story" in the context of the timeline of the video. For a selected element in (5), (6) presents shot thumbnails from that element. (4) is a video player component that also presents a preview image of a shot or story when the user moves the mouse over the corresponding shot or story. (7) shows a collection of user selected shot thumbnails. In their interface they use *visual clues* to indicate which content item has been previously visited or explicitly excluded from search. Moreover, their system allows performing a multiple user collaborative search.

Story-Based Video Retrieval Interfaces

The systems which have been introduced in the previous section support users in retrieving *shots* of a video. While this approach is useful in some cases, shots are not the ideal choice in other cases. In this section, we survey systems that provide users access to *news stories*.

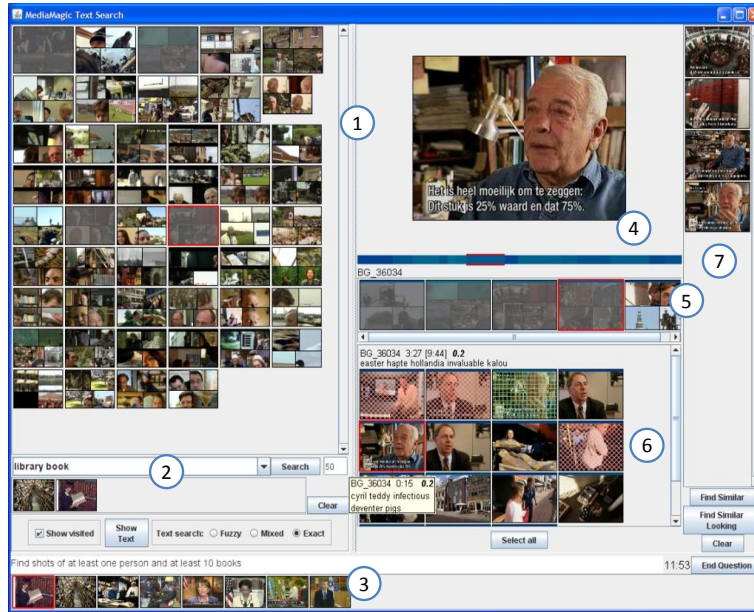


FIGURE 2.8: The *MediaMagic* video search interface [Adcock et al. \[2008\]](#)

[Pickering et al. \[2003\]](#) introduce a web-based video retrieval interface that allows access to latest news video stories. The interface of their ANSES system, shown in Figure 2.9, supports full text search, retrieval results are ranked accordingly. Each retrieved story is represented by the first key frame of the news story, descriptive meta-data (broadcasting date and length of the video) and the story transcript is displayed. Further, key entities, namely organisations, persons, locations and dates are visualised using different colour codes. Users can access the full story by clicking on a provided link.

[Haggerty et al. \[2004\]](#) introduce NewsFlash. As shown in Figure 2.10, their system supports two forms of search: Full text search and profile search. Full text search is triggered in the query formulation panel. Profile search is performed in the profile search panel. Each retrieved result is visualised by a representative key frame. Moving the mouse over a key frame highlights a query biased summary of the story transcript. A click on a result starts playing the video in the top right corner of the interface. A similar system is introduced by [Morrison and Jose \[2004\]](#).

Figure 2.11 shows a screenshot of Físchlár-News, an online news video archive introduced by [Lee et al. \[2006\]](#). Their web interface support query-by-textual-keyword. Retrieval results are displayed in descending order of relevance to the given query. Each result is represented by a key frame, descriptive metadata and a summary of the story transcript. Clicking on “Play this story” starts streaming the corresponding video. Further, users can provide explicit relevance feedback, thus creating a profile. This profile

2.1. Basic Concepts of Video Retrieval



FIGURE 2.9: The ANSES video search interface [Pickering et al., 2003]



FIGURE 2.10: The NewsFlash video search interface [Haggerty et al., 2004]

can be exploited by clicking on the button labelled “Recommended” on top of the interface. This action triggers a personalised search query. On the left hand side of the interface, a calendar is displayed, which allows users to select all stories of a specific date.

2.1. Basic Concepts of Video Retrieval



FIGURE 2.11: The *Físchlár-News* video search interface [Lee et al., 2006]

Diriye et al. [2010] introduce NewsRoom, a web-based video retrieval interface that provides access to up-to-date news video content. Similar to the previous interfaces, it supports full text search, retrieval results are ranked based on their relevance. News stories are represented by a representative key frame, descriptive metadata (broadcasting date and video story length) and a summary of the transcript, determined using lexical chains. Further, moving the mouse over a retrieved result, additional information is displayed, i.e. related video and the topic of the story. Clicking on a result will start streaming the video from the given time point of the news story. After triggering a textual retrieval, NewsRoom identifies *information landmarks*, that shall support the user in understanding the overall direction of the retrieved results. As shown in Figure 2.12, the topics of the search results, determined by performing a textual analysis of the retrieval results are visualised on top of the interface. Salient topics are emphasised. Another feature, visualised on the left hand side of the interface are news categories. It allows users to focus their information seeking task, i.e. by narrowing down the displayed topics.

Focusing on exploiting semantic knowledge, Bürger et al. [2005] introduce the Smart Content Factory, an infrastructure aiming at providing a “semantic layer” on top of news broadcast. Each story in the index has been enriched with semantic information, i.e. places mentioned in the transcript are matched with a generic geography thesaurus. Further, the topic of each story is determined based on key words in the transcript. The interface of their system, shown in Figure 2.13, supports query-by-textual-keyword. Retrieval results are visualised by a representative key frame, extracted places and topics and a summary of the text transcript. Each mentioned place in the transcript is linked with a hyperbolic tree, where the place is shown in its geographic context. Users can

2.1. Basic Concepts of Video Retrieval

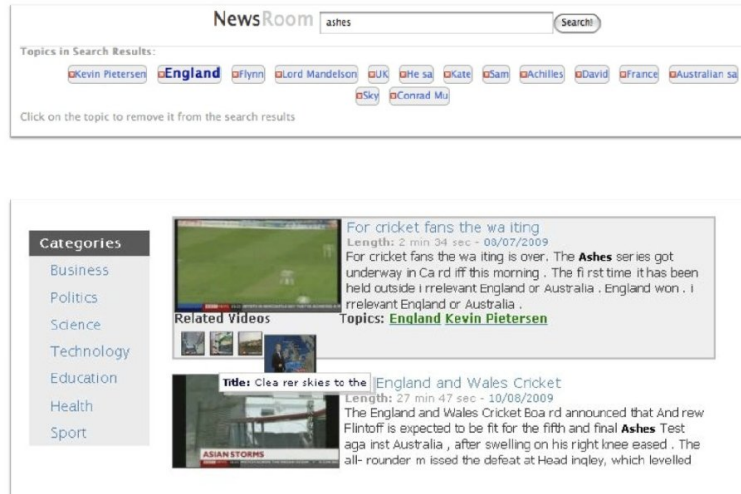


FIGURE 2.12: The *NewsRoom* video search interface [Diriye et al., 2010]

interact with this tree, hence browsing the collection based on the semantic content of each story.

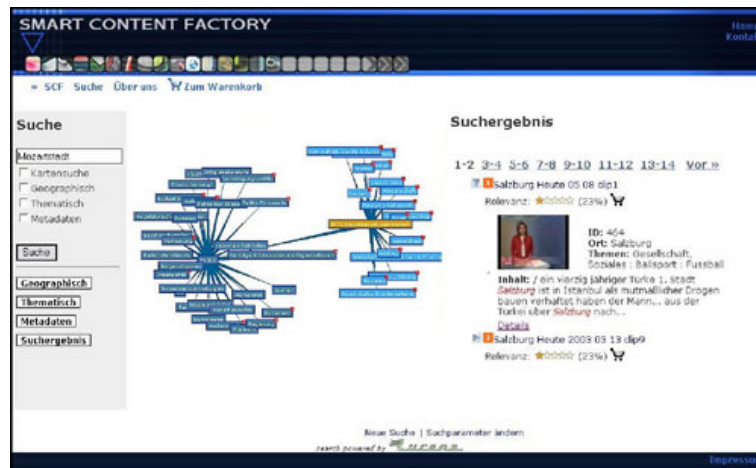


FIGURE 2.13: The *Smart Content Factory* video search interface [Bürger et al., 2005]

Discussion

In this section, we surveyed state-of-the-art video retrieval interfaces that support two different units of retrieval. The majority of interfaces that have been developed treat the video *shot* as unit of retrieval. As we have seen, the de-facto standard to represent video shots is by extracting representative key frames. Even though different in nature, most interface consist of two main panels: A querying panel and a result list panel. Further, most interfaces list relevant video shots in descending relevance to the given

query. The second type of interfaces treat *news stories* as basic unit of retrieval. As we have shown, far less system implementations exist for this scenario. The main reason for this imbalance is certainly the influence of the TRECVID evaluation campaign on video retrieval research: Most interfaces have been evaluated using the TRECVID data collection where the unit of retrieval is the video shot.

Within this thesis, we will evaluate different research hypotheses that require graphical user interfaces. We will therefore introduce two different interfaces. While the first interface is designed to retrieve and recommend video shots, the second interface visualises news videos.

2.1.6 Summary

In this section, we surveyed basic principles of video retrieval. We first introduced the general structure of video documents. As we have shown, videos consist of audio-visual data streams and are often accompanied with metadata. Metadata is mainly of textual nature. Further, we argued that most retrieval scenarios require videos to be split into smaller units of retrieval. We surveyed two different segmentation units and their application: video shots and (news) video stories. Moreover, we introduced document representation and matching techniques, including indexing, retrieving and ranking of video documents. Finally, we introduced different graphical user interfaces that support users in their information seeking task.

The techniques and methods we have introduced combine all required parts of a video IR system. In the next section, we will survey how these approaches can be applied to adaptively support users' information needs.

2.2 Personalised Video Search and Recommendation

This section surveys personalised video search and recommendation. In Section 2.2.1, we provide an overview on the research area. One challenge toward personalisation is capturing user interests, which is surveyed in Section 2.2.2. Another challenge is to interpret this interest, hence to provide personalisation services. Section 2.2.3 introduces existing personalisation techniques. A state-of-the-art survey on personalisation and recommendation approaches is given in Section 2.2.4. Section 2.2.5 summarises and concludes this section. Unless stated otherwise, this section is based on [Brusilovsky et al., 2007; Belew and van Rijsbergen, 2000].

2.2.1 Personalisation Overview

In the previous section, we surveyed basic concepts of video retrieval. We argued that when interacting with a video retrieval system, users express their information need in search queries. The underlying retrieval engine then retrieves relevant retrieval results to the given queries. A necessary requisite for this IR scenario is to correctly interpret the users' information need. As [Spink et al. \[1998\]](#) indicate though, users very often are not sure about their information need. One problem they face is that they are often unfamiliar with the data collection, thus they do not know what information they can expect from the corpus [[Salton and Buckley, 1997](#)]. Further, [Jansen et al. \[2000a\]](#) have shown that video search queries are rather short, usually consisting of approximately three terms. Considering these observations, it is hence challenging to satisfy users' information needs, especially when dealing with ambiguous queries. Triggering the short search query "Victoria", for example, a user might be interested in videos about cities called Victoria (e.g. in Canada, United States or Malta), landmarks (e.g. Victoria Park in Glasgow or London), famous persons (e.g. Queen Victoria or Victoria Beckham) or other entities called Victoria. Without further knowledge, it is a demanding task to understand the users' intentions. Interactive information retrieval aims at improving the classic information retrieval model that we introduced in the previous section by studying how to further engage users in the retrieval process, in a way that the system can have a more complete understanding of their information need. Thus, aiming to minimise the users' efforts to fulfil their information seeking task, there is a need to personalise search. In a web search scenario, [Mobasher et al. \[2000\]](#) define personalisation as "any action that tailors the Web experience to a particular user, or a set of users". Another popular name is adaptive information retrieval, which was coined by [Belew \[1989\]](#) to describe the approach of adapting, over time, retrieval results based on users' interests. In this thesis, we will use both terms synonymously.

In this section, we will introduce basic concepts of personalised IR. In Section [2.2.2](#), we first survey different sources that are used to gather users' interests, an important requisite for any type of personalisation. Section [2.2.3](#) introduces different application techniques. In Section [2.2.4](#), we introduce state-of-the-art personalisation approaches with a main focus on the video domain. Section [2.2.5](#) summarises the personalisation section.

2.2.2 Gathering and Representing Interest

Most of the approaches that follow the interactive information retrieval model are based on relevance feedback techniques [[Salton and Buckley, 1997](#)]. Relevance feedback is

one of the most important techniques within the IR community. An overview of the large amount of research focusing on exploiting relevance feedback is given by [Ruthven and Lalmas \[2003\]](#). The principle of relevance feedback is to identify the user's information need and then, exploiting this knowledge, adapting search results. [Rocchio \[1971\]](#) define relevance feedback as follows: The retrieval system displays search results, users provide feedback by specifying keywords or judging the relevance of retrieved documents and the system updates the results by incorporating this feedback. The main benefit of this approach is that it simplifies the information seeking process, e.g. by releasing the user from manually reformulating the search query, which might be problematic especially when the user is not exactly sure what they are looking for or does not know how to formulate their information need. There are three types of relevance feedback in interactive information retrieval which will be introduced in the remainder of this section: explicit, implicit and pseudo relevance feedback. Further, we introduce the Ostensive Model of Developing Information Need, that addresses the problem of non-static interest and discuss approaches to represent user interests in personal profiles.

Explicit Relevance Feedback

A simple approach to identify users' interests is to explicitly ask them about their opinion. In a retrieval context, they can express their opinion by providing *explicit relevance feedback*. Hence, the user is asked during their retrieval process to actively indicate which documents are relevant in the result set. This relevance judgement can be given on a binary or graded relevance scale. A binary feedback indicates that the rated document is either relevant or non-relevant for the user's current information need. Considering that binary relevance requires a rather strong judgement, a relevance scale allows the user to define different grades of relevance such as "highly relevant", "relevant", "maybe relevant" or "somewhat relevant". As of May 2010, the commercial video sharing platform YouTube supports binary feedback by providing a "Thumbs up" button in their interface. The Dailymotion platform opted for the graded relevance scale scheme. Registered users of their service can express their interest on a Five "star" scale. Explicit relevance feedback is very reliable. Although the impact of explicit relevance feedback in above systems remains unclear, it has been shown in text retrieval that giving explicit relevance feedback is a cognitively demanding task and can affect the search process. Also, previous evaluations have found that users of explicit feedback systems often do not provide sufficient levels of feedback in order for adaptive retrieval techniques to work [[Hancock-Beaulieu and Walker, 1992](#); [Belkin et al., 2001](#)].

Implicit Relevance Feedback

Deviating from the method of *explicitly* asking the user to rate results, the use of *implicit* feedback techniques helps learning users' interest unobtrusively. The main advantage is that this approach delivers the user from providing explicit feedback. As a large quantity of implicit data can be gathered without disturbing the users' workflow, the implicit approach is an attractive alternative. According to [Nichols \[1998\]](#), however, information gathered using implicit techniques are less accurate than information based on explicit feedback. [Agichtein et al. \[2006\]](#) evaluated the effect of user feedback on web retrieval using over 3000 queries and 12 million user interactions. They show that implicit relevance feedback can improve retrieval performance by much as 31% relative to systems that do not incorporate any feedback. Furthermore, both implicit and explicit measures can be combined to provide an accurate representation of the users' interests. [Kelly and Teevan \[2003\]](#) provide a literature overview of the research which has been done in the field.

Not all implicit measures are useful to infer relevance. Thus, various research has been done to detect those features which promise to be valid indicators of interest.

From the psychological point of view, a promising indicator of interest is viewing time. People look at objects or things they find interesting for a longer time than on uninteresting things. For instance, [Faw and Nunnally \[1967\]](#) showed a positive correlation between "pleasant ratings" and viewing time and [Day \[1966\]](#) found that most people look longer on images they liked than on images they disliked. According to [Oostendorp and Berlyne \[1978\]](#), people look longer at objects evoking pleasurable emotions. Transferring these findings into an information retrieval context, users are expected to spend more time in viewing relevant documents than non-relevant documents. [Claypool et al. \[2001\]](#) introduce a categorisation of both explicit and implicit interest indicators in web retrieval. They conclude that time spend on a page, the amount of scrolling on a page and the combination of these two features are valid implicit indicators for interest. Furthermore, they found that individual scrolling measures and the number of mouse clicks are ineffective indicators. [Morita and Shinoda \[1994\]](#) evaluated if user behaviour while reading newsgroup articles could be used as implicit indicator for interest. They measured the copying, saving or following-up of an entry and the time spend for reading the entries. They reveal that the reading time for documents rated as interesting was longer than for uninteresting documents. A relation between interest and following-up, saving or copying was not found. [White et al. \[2002\]](#) consider reading time as a technique to automatically re-rank sentence-based summaries. Their results, however, were inconclusive. [Kelly \[2004\]](#) criticises the study approaches that focus on display time

as relevance indicator, as she assumes that information-seeking behaviour is not influenced by contextual factors such as topic, task and collection. Therefore, she performed a study to investigate the relationship between information-seeking task and the display time. Her results cast doubt on the straightforward interpretation of dwell time as an indicator of interest or relevance.

Another indicator of interest which has been analysed is the users' browsing behaviour. [Seo and Zhang \[2000\]](#) introduce a method to learn users' preferences from inobtrusively observing their web-browsing behaviour. They conclude that their approach can improve retrieval performance. However, the adaptation of users' interest over a longer period of time has not been taken into account as their search sessions were set up for a short period only. [Maglio et al. \[2000\]](#) suggest to infer attention from observing the eye movements. In the HCI community, this has become a common technique.

Pseudo Relevance Feedback

A third relevance feedback approach is called *pseudo, blind* or *ad-hoc relevance feedback*. It was first introduced by [Croft and Harper \[1997\]](#). Differing from the previous two approaches, pseudo relevance feedback does not require users providing relevance assessments; the top ranked retrieval results are considered being relevant and used to adapt the initial search query. Considering the lack of manual input, its usage as source for personalisation techniques is questionable.

Evolving User Interest

In a retrieval context, profiles can be used to contextualise the user's search queries within his or her interests and to rerank retrieval results. This approach is based on the assumption that the user's information interest is static, which is, however, not appropriate in a retrieval context. [Campbell \[1995\]](#) argues that the user's information need can change within different retrieval sessions and sometimes even within the same session. He states that the user's search direction is directly influenced by the documents retrieved. The following example illustrates this observation:

Imagine a user who is interested in red cars and uses a video retrieval system to find videos depicting such cars. Their first search query returns several video clips, including videos of red Ferraris. Watching these video, he or she wants to find more Ferraris and adapts the search query accordingly. The new result list now consists of video clips showing red and green

Ferraris. Fascinated by the rare colour for this type of car, he/she again reformulates the search query to find more green Ferraris. Within one session, the user's information need evolved from red cars to green Ferraris.

Based on this observation, [Campbell and van Rijsbergen \[1996\]](#) introduce the Ostensive Model of Developing Information Need that incorporates this change of interest by considering, *when* a user provided relevance feedback. In the Ostensive Model, providing feedback on a document is seen as ostensive evidence that this document is relevant for the user's current interest. The combination of this feedback over several search iterations provides ostensive evidence about the user's changing interest. The model considers the user's changing focus of interest by granting the most recent feedback a higher impact over the combined evidence.

User Profiling

Considering the large amount of personal data that can be captured, most personalisation systems rely on user profiles to manage this data. User profiling is the process of learning user interests over an extended period of time. User profiles may contain demographic information and user feedback, that they expressed either explicitly or implicitly. In this section, we will highlight basic principles of user profiling. A state-of-the-art survey is given by [Gauch et al. \[2007\]](#), who distinguish between two types of user profiles: *short-term* and *long-term* profiles. Short-term user profiles are used for personalisation within one session, i.e. any feedback that the user provides during their current information seeking task is used to adapt the results. Long-term user profiles, on the other hand, aim to keep track of users' long-term interests. Personalisation services based upon such profiles can hence adapt results considering user feedback which was given over multiple sessions.

In literature, three types of user profile representations exist: Weighted keywords or concepts, semantic networks and association rules. Association rules are mainly applied in the field of web log mining. By identifying relations between variables, it is possible to identify popular variable combinations. Association rules rely on large amount of data, often provided by different users. Considering this requirement, we therefore focus on weighting-based profiles and semantic network-based profiles, neglecting association rules. A survey on association rules is given by [Mobasher \[2007\]](#).

Weighted Keywords or Concepts The most popular representation of user interests is the weighted keyword or concept approach. Interests are represented as a vector of weighted terms that have either been extracted from those documents that users showed

interest in or that have been provided manually by the users. The weighting indicates the importance of the corresponding term in the user profile. The main disadvantage of this approach is the so-called polysemy problem, hence the multiple meanings that each word can have.

An early example includes the *Amalthaea* system [Moukas and Maes, 1998] where keywords, extracted from websites are assigned with a weighting based on TF.IDF [Baeza-Yates and Ribeiro-Neto, 1999]. The terms-weighting combination is represented in the profile as a vector. Similar approaches are studied by Sakagami and Kamba [1997], who introduce personalised online newspapers, Lieberman [1995], introducing a browsing assistant and Pazzani et al. [1996], who propose a recommender system that exploits weighted keyword profiles.

Even though weighted keyword profiling has been well studied in the text domain, hardly any work has been done on studying similar approaches in the video domain. Few exceptions include Weiß et al. [2008], who, however, generate user profiles exploiting video metadata rather than audio-visual features.

Semantic Networks In the semantic network approach, keywords are replaced with concepts. User interests are represented as weighted nodes of a graph where each node is a concept in which the user showed interest in. A similar approach is referred to as concept profiles. Differing from semantic networks, however, concept profiles consider abstract topics rather than specific words to represent user interests.

The advantage of semantic networks and concept profiles is that concepts can be organised in a hierarchical structure. In a web search scenario, for example, a user might have shown interest in a website that has been categorised in the Open Directory Project (ODP)²⁻³ as being a website about “Travel and Tourism in Scotland”. Within ODP, “Travel and Tourism in Scotland” is a subcategory of “Travel and Tourism in the United Kingdom”. Bloedorn et al. [1996] argue to exploit such hierarchies, since they allow generalisation of concepts. Personalisation services could hence exploit this relationship, e.g. by recommending other websites belonging to the same or more general categories. Various approaches have been studied to exploit such public hierarchies:

Daoud et al. [2008, 2009] analyse the documents that users’ provided implicit relevance feedback on, map the concepts of these documents with the Open Directory Project (ODP) ontology and store them in a hierarchical, graph-based user profile at the end of each search session. Other personalisation techniques based on ODP include [Chirita et al., 2005; Sieg et al., 2007; Chaffee and Gauch, 2000; Tanudjaja and

²⁻³<http://dmoz.org/> Last time accessed on: 5 May 2010. The ODP is a manually edited catalog of websites. It is organised as a tree, where websites are leaf nodes and categories are internal nodes.

Mui, 2002], who show that incorporating this taxonomy can significantly outperform unpersonalised search techniques.

Dudev et al. [2008] propose the creation of user profiles by creating knowledge graphs that model the relationship between different concepts in the Linked Open Data Cloud²⁻⁴. Different from the ODP ontology, important parts of the Linked Open Data cloud have been created automatically, e.g. by converting Wikipedia pages into an ontological representation. Consequently, the available data is of immense size, but rather un-uniform.

Gauch et al. [2007] argue that when exploiting such public directories, various pre-processing steps have to be performed, including transforming the content into a concept hierarchy or dealing with situations where some concepts might have multiple entries while other concepts are less important. Moreover, they argue that the more levels of the hierarchy are used, the more general the profile representation might become.

Discussion

In this section, we surveyed issues regarding gathering user interests. We first introduced relevance feedback, the most common technique used to identify user interest. As we have shown, relevance feedback techniques can be split into three main categories: explicit, implicit and pseudo relevance feedback. Rui et al. [1998b] propose interactive relevance feedback as a method to bridge the Semantic Gap, assuming that high-level concepts can be identified using low-level features. In their approach, users have to rate images according to their relevance for the information need. The features are weighted automatically to model high-level concepts based on user's feedback. The results of their study promise a reduction of query formulation efforts, as the relevance feedback technique seems to gather the user's information need effectively. According to Huang and Zhou [2001], relevance feedback appears to be more promising in the image field than in the classical text field, as it is easier and faster to rank images according to their relevance than ranking text documents. Various research has been done to optimise parameters and algorithms [Zhou and Huang, 2001; Doulamis and Doulamis, 2003; Huang and Zhou, 2001; Porkaew and Chakrabarti, 1999]. Karthik and Jawahar [2006] introduce a framework to evaluate different relevance feedback algorithms. They conclude that statistical models are the most promising algorithms. These models try to cluster images into relevant and non-relevant images.

In this thesis, we aim to evaluate the use of relevance feedback to improve video retrieval. We will focus on implicit relevance feedback, aiming to evaluate the perfor-

²⁻⁴<http://linkeddata.org/> Last time accessed on: 5 May 2010. The Linked Open Data collection of ontologies unites information about many different freely available concepts, ODP being one of them.

mance bounds of this information source.

Further, we introduced the Ostensive Model of Developing Information Need, which emphasises the time when users provided relevance feedback. Various forms of this model have been developed and applied in image retrieval [Campbell, 2000; Urban et al., 2006b; Leelanupab et al., 2009b] and Web search scenarios [Joho et al., 2007]. Within this thesis, we aim to study its usability in video personalisation to model user interests over a longer time period.

Finally, we introduced different approaches of user profiling. User profiling is one of the key challenges in adaptive search and recommendation. As we discussed, two types of user profiling exist: short-term and long-term profiling. Within this thesis, we will employ both approaches to study the use of implicit relevance feedback in the video domain.

2.2.3 Personalisation Techniques

After introducing approaches to gather users' interests, this section introduces personalisation techniques that exploit this feedback. Jameson [2008] lists various adaptation paradigms, including ability-based user interfaces, learning personal assistants, recommender systems, adaptation to situational impairments, personalised search and user interfaces that adapt to the current task. Note that within this work, we will focus on recommender systems and personalised search, neglecting the other paradigms. Both paradigms will be introduced in the remainder of this section.

Personalised Search

In 2008, Marissa Mayer, the Vice President of Search and User Experience of Google Inc. predicted in an interview held at the LeWeb conference that “in the future personalised search will be one of the traits of leading search engines” [Mayer, 2008]. This statement reflects the increasing attention that personalised search draws from both academia and industry sides. Teevan et al. [2010] argue that there is a big performance difference between personalised and non-personalised search engines. They hence argue that there is a big potential for personalisation.

As discussed before, one of the main problems in IR is that users express their information need using short queries only. Matching these short queries with the document collection will return only a relative small amount of relevant results. Providing more search terms could improve the retrieval performance, as then, more documents can be retrieved. Dependent on how much influence the user shall have, the expansion terms can either be added by the system – *automatic query expansion* or by the user – *inter-*

active query expansion. According to Ruthven [2003], query expansion terms which have been provided in an interactive process are less useful than automatically identified term. We therefore neglect interactive query expansion and focus on automatic query expansion techniques.

Automatic query expansion based on relevance feedback is a common technique to refine search queries (e.g. [Rocchio, 1971; Salton and Buckley, 1997]). The reason for its success can be found by the users: they tend *not* to give relevance feedback or to formulate their search queries appropriate. The first query expansion approach was introduced by Rocchio [1971]. In an retrieval experiment, Rocchio added term vectors for all retrieved relevant documents and subtracted the term vectors for all irrelevant documents to refine search queries. Hence, terms are aligned with a weighting which can increase and decrease during the process. Järvelin et al. [2001] has shown that concept-based query expansion, i.e. exploiting ontologies, is helpful to improve retrieval performance. Multiple other studies show the effectiveness of ontology-based expansion [Bhagal et al., 2007].

Video retrieval based query expansion approaches include [Volkmer and Natsev, 2006], who rely on textual annotation (video transcripts) to expand search queries. Within their experiment, they significantly outperform a baseline run without any query expansion, hence indicating the potentials of query modification in video search. Similar results are reported by Porkaew and Chakrabarti [1999] and Zhai et al. [2006], who both expand search queries using content-based visual features.

Document Recommendation

Another personalisation technique is document recommendation. Anderson [2006], editor-in-chief of the Wired Magazine, claims that “we are leaving the Information Age and entering the Recommendation Age.” Differing from personalised search, where search results are adapted based on user interests, recommender systems provide additional items (documents). The main idea of this technique is to provide users faster and more information they might be interested in. Further, in an e-commerce scenario, recommendations shall influence the users. For example, Amazon.com²⁻⁵ provides recommendations on other products that their customers might be interested in. Recommender systems hence inform users about things they might not be aware of and have not been actively searching for. They can be distinguished into two main categories: *content-based recommender systems* and *collaborative filtering systems*. In the remainder of this section, we will briefly introduce both approaches and discuss user profiling

²⁻⁵<http://www.amazon.com/> last time accessed: 5 May 2010

issues.

Content-Based Recommender Systems Content-based recommender systems determine the relevance of an item (e.g. a video document, website or a product) based on the user's interest in other, similar items. Items in the data collection are evaluated in accordance to users' previous feedback and the most similar items are recommended to the user. A survey is given by [Pazzani and Billsus \[2007\]](#). User interest is usually stored in personal user profiles.

Collaborative Filtering Collaborative filtering systems aim to exploit the opinion of people with similar interests. Thus, items are recommended when other users of the recommender system showed interest in it. Differing from content-based recommendation, where the content of the documents has to be analysed, the challenge in collaborative filtering is in identifying users with similar interests. A survey is given by [Schafer et al. \[2007\]](#).

Discussion

In this section, we introduced two types of personalisation techniques: Personalised search and document recommendation. Even though both techniques can rely on the same information, their application differs. While personalised search aims to adapt retrieval results based on the users' previous interests, recommender systems provide additional information that users did not necessarily ask for. In this thesis, we employ both personalisation techniques.

2.2.4 State-of-the-art Personalisation Approaches

Implicit feedback techniques have been successfully applied to retrieval systems in the past. For instance, [White \[2004\]](#) and [Joachims et al. \[2005\]](#) defined and evaluated several implicit feedback models on a text-based retrieval system. Their results indicated that their implicit models were able to obtain a comparable performance to that obtained with explicit feedback models. However, their techniques were based on textual information, and applied individually at runtime during the user's search session. As stated previously, the lack of textual annotation on video digital libraries prevents the adoption of this approach in video retrieval systems. One solution to this problem is to collect the implicit information from a large number of past users, following a collaborative recommendation strategy.

The exploitation of usage information from a community of past users is a widely researched approach to improve information retrieval systems. However, most of past and present studies focus on the text retrieval domain. The main hypothesis of such systems is that when a user enters a query, the system can exploit the behaviour of past users that were performing a similar task. For instance, [Bauer and Leake \[2001\]](#) build up a task representation based on the user's sequence of accessed documents. This task representation is used by an information agent, which proactively suggests documents to the user.

A commonly exploited past usage information structure is clickthrough data. Clickthrough data is limited to the query that the user executed into the system, the returned documents, and the subsequent documents that the user opened to view. [Sun et al. \[2005\]](#) and [Dou et al. \[2007\]](#) mine query log clickthrough information to perform a collaborative personalisation of search results, giving preference to documents that similar users had clicked previously for similar queries. [Sun et al. \[2005\]](#) apply a dimensional reduction pre-processing step on the clickthrough data in order to find latent semantic links between users, queries and documents. [Dou et al. \[2007\]](#) complement these latent relationships with user-topic and document-topic similarity measures. [Craswell and Szummer \[2007\]](#) use a bipartite graph to represent the clickthrough data of an image retrieval system, where queries and documents are the nodes and links are the ones directly captured in clickthrough data. A random walk is then applied in order to recommend images based on the user's last query.

Following this line of works, [White et al. \[2007\]](#) introduced the concept of query and search session trails, where the interaction between the user and the retrieval system is seen as a trail that leads from the user query to the last accessed document of the query session or the search session, which is constituted by multiple query sessions. They argue that the user's need of information was most likely satisfied with the last document of these trails, i.e. the last document that the user accessed in the query or search session.

Above approaches rely on text to base personalisation techniques on. As we discussed, however, multimedia documents consist of multiple modalities, including text and audio-visual features. A comprehensive survey on relevance feedback in the image domain is given by [Zhou and Huang \[2003\]](#). The main problem when incorporating content-based features is to find out which feature represents the image best. Further, which approach should be followed to build an adaptive retrieval model. This problem even increases when dealing with video content, since additional features can be considered.

An early work focusing on relevance feedback in the video domain is presented by

[Yong et al. \[1997\]](#). They illustrate that the content rich nature of multimedia document require a more precise feedback. When a user provides relevance feedback on a video document, it is not clear, which feature of this document should be exploited to identify similar documents. It could be, for example, the colour feature of the video or the context of the video that is relevant to the user. In their work, they propose a relevance feedback framework for multimedia databases where search queries are adapted based on user's relevance feedback. Their framework, however, is focusing on explicit relevance feedback only, neglecting the possibility to exploit implicit indicators of relevance. Another study focusing on explicit relevance feedback in the video domain is provided by [Hauptmann et al. \[2006\]](#), who asks users to manually provide labels for video shots. After retrieving similar video documents for the label, users are then asked to explicitly mark those videos that match to the corresponding label. [Doulamis et al. \[1999\]](#) requires explicit relevance feedback, given during the information seeking process, to update video retrieval results. Hence, in all three studies which have been mentioned above, explicit relevance feedback is used to personalise search queries.

[Luan et al. \[2007\]](#) consider text, high-level features and multiple other modalities to base their relevance feedback algorithms on. In their approach, however, they do not focus on personalising techniques but rather on improving video annotation. An extension of their work is presented in [Luan et al. \[2008\]](#), where they propose multiple feedback strategies that support interactive video retrieval. They argue that the more complex nature of video data, when compared with textual data, require different types of feedback strategies. Therefore, they distinguish between three categories of feedback types: recall-driven, precision-driven and temporal locality-driven feedback. The first type focuses on analysing the correlation of video features that can be extracted from positive and negative rated video documents. It aims to result in higher recall values. The second type employs active learning techniques and aims to constantly re-rank retrieval results. Its purpose is to increase precision. The third type, temporal locality-driven feedback, exploits the temporal coherence among neighboured shots. In their study, they show that giving the user the choice to chose between these different feedback types can effectively improve video retrieval performance.

[Yang et al. \[2007\]](#) study video recommendations based on multimodal fusion and relevance feedback. They exploit viewing time to identify positive recommendations in their video recommender system. They interpret a very short video viewing time as negative relevance feedback, arguing that the user is not interested in the video's content. Further, they argue that a longer viewing time indicates a higher interest in the video. Another study focusing on negative relevance feedback is introduced by [Yan et al. \[2003\]](#), who consider the *lowest* ranked documents of a search query to be not

relevant. Hence, they suggest to re-formulate the initial search query by using these documents as negative example. Their approach, however, does not require any specific user input, it is based on pseudo-relevance feedback only. Nevertheless, their findings indicate that pseudo-relevance feedback can successfully be employed to adapt retrieval results.

[Vrochidis et al. \[2010\]](#) aim to improve interactive video retrieval by incorporating additional implicit indicators of relevance. They consider the following user actions as implicit relevance feedback: (1) Text-based search queries. Given a search query, they assume that the keywords that form the search query are relevant. (2) Visual queries. If a user provides a key frame as visual query, they assume this key frame to be relevant. (3) Side-shot and video-shot queries. When a user performs an action on a shot, they assume this shot to be relevant. Arguing that each indicator can be interpreted to a different degree to be relevant, they suggest to assign each feedback type a different weighting. Within this thesis, we study implicit relevance feedback on the same lines.

A purely content-based personalisation approach is introduced by [Aksoy and Çavuş \[2005\]](#) who extract low-level visual from explicitly relevant rated key frames. They suggest to extract different feature vectors from those key frames and to assign a relevance weighting for each vector.

A different understanding of implicit relevance feedback is introduced by [Villa et al. \[2008a\]](#). Within their study, they aimed to elaborate whether the awareness of other users' search activities can help users in their information seeking task. They introduce a search scenario where two users remotely search for the same topic at the same time. Using their interface, each user is able to observe the other user's search activities. They conclude that awareness can have a direct influence in the information seeking process, since users can learn from other users' search results and/or search queries. In case users copy the other user's search activity, this action can be interpreted as implicit relevance feedback, that the action has been performed on (or resulted in) relevant documents. A similar study is performed by [Halvey et al. \[2009b\]](#), who, however, study the impact of awareness in an asynchronous scenario. Within their experiment, users can interact with other user's previous search sessions. Again, continuing other user's search sessions can be seen as implicit indication that the retrieved search results are partially relevant.

Discussion

In this section, we surveyed various state-of-the-art personalisation approaches. As we have shown, the most recent research approaches exploit users' clickthrough data to

2.3. Evaluation Methodologies

provide either recommendations or to personalise search. Hence, implicit relevance feedback is employed to determine user interests. Further, to best of our knowledge, hardly anything has been done to incorporate implicit relevance feedback in the video retrieval and recommendation domain. In this thesis, we hence aim to study the use of implicit indicators of relevance. Therefore, we adopt the introduced concept of search trails in one of our recommendation models. While White et al. exploit only the last documents of the search trails, we are interested in representing and exploiting the whole interaction process, based on the hypothesis that in video retrieval this continuous path contains relevant evidence than can be exploited to achieve better performance in collaborative recommendation.

2.2.5 Summary

In this section, we surveyed personalised video search and recommendation. An important prerequisite for any kind of personalisation service is to identify users' personal interests. As we have shown, the most popular technique to gather this interest is relevance feedback. After introducing different feedback techniques and challenges such as evolving interest and user profiling, we introduced different personalisation services, namely personalised search and recommender systems. Finally, we surveyed state-of-the-art personalisation and recommendation systems.

2.3 Evaluation Methodologies

This section surveys well-established evaluation methodologies in the information retrieval domain. An overview of information retrieval evaluation is provided in Section 2.3.1. The most commonly used evaluation measures are introduced in Section 2.3.2. Three main evaluation approaches are dominating the research field which are introduced in Sections 2.3.2, 2.3.3 and 2.3.4, respectively. Section 2.3.5 summarises and concludes this section. Unless stated otherwise, the material in this section is based on [Campbell and Stanley, 1963; Ingwersen and Järvelin, 2005; Voorhees and Harman, 2005; Dasdan et al., 2010].

2.3.1 Evaluation Overview

The standard process of scientific research is to evaluate hypotheses and research questions based on clear and justified standards. In the IR community, evaluation has a long tradition, mainly due to the implementation of the Text REtrieval Conference (TREC)

initiative [Voorhees and Harman, 2005]. TREC, organised by the (US American) National Institute of Standards and Technology (NIST) supports research in information retrieval by providing the necessary infrastructure for large-scale evaluation of retrieval methodologies. IR systems, approaches and methodologies are usually evaluated in accordance to their effectiveness and computational efficiency. System effectiveness can be deducted by analysing two features: The system’s ability to model relevance, i.e. to correctly associate documents to a given query and its ability to present these results on a graphical user interface. The majority of IR experiments focus on evaluating the system effectiveness. In this system-centred evaluation scheme, the system effectiveness is evaluated, using well-established evaluation measures, by analysing the system’s output with respect to (mostly) manually created relevance assessments lists. We will introduce this scheme in Section 2.3.2. In order to evaluate the presentation of the results, different interface models or their usability, a user-centred evaluation scheme is employed. This scheme, borrowed from research on human-computer interaction relies on questionnaires and usage log analyses to evaluate the system effectiveness. Section 2.3.3 surveys user-centred evaluation of IR systems. A third evaluation scheme, simulation-based evaluation aims to combine the best of both worlds; the advantage of batch evaluation and using standard evaluation measures as well as user’s interactions. It is often employed when user interaction would be required to evaluate user-centred research approaches, but the implementation of a user-centred study is not appropriate. We survey simulation-based evaluation in Section 2.3.4. Section 2.3.5 summarises the evaluation survey section.

2.3.2 System-centred Evaluation

The most common evaluation scheme in the IR community is system-centred evaluation [Ingwersen and Järvelin, 2005]. Its success is due to the well-established evaluation methodology that is mainly promoted by TREC, the premium evaluation platform within the community. System-centred evaluation is based on early work of Cleverdon et al. [1966], who introduced a test dataset in a controlled setting for the evaluation of computer-based retrieval engines, often referred to as *Cranfield Paradigm*. In their work, they performed various retrieval experiments on different test databases in a controlled environment. Constraining the dataset helped them to identify available relevant documents which is helpful in drawing a conclusion on the quality of the output of a retrieval engine. Cleverdon [1970] conducted further experiments with alternative indexing languages constituting the performance variables under investigations. These experiments are known as *Cranfield II*. Two assumptions underlie the methodology:

First of all, users only want to retrieve results which are relevant to their query and are not interested in non-relevant results. Furthermore, the relevance of a document to a query is uniform to all. We will discuss this point in more detail later. The setting of a classical Cranfield experiment can be divided into three components:

1. A static *test corpus* of documents. The purpose of test collections is to provide common test beds that enable comparison of novel and existing research approaches. Within the last few years, various test collections have been introduced to promote research in different retrieval research domains. Example are Web Test collections²⁻⁶, image collections²⁻⁷ or Blog²⁻⁸ collections. Due to the leading role of TREC, we will further refer to such standardised test collections as TREC-collections.
2. A *set of queries* that are created based on the content of the documents of the test collection. The queries serve, together with the collection, as input for the retrieval engine.
3. A set of documents judged to be relevant or non-relevant to each query (*relevance assessments*), also referred to as qrels. Retrieval results for each query will be compared to these judged documents to pose a statement about the performance of the retrieval engine. Saracevic [1996] distinguishes between five types of relevance: topical, cognitive, motivational, system and situational relevance. A thorough discussion about these different types is given by Borlund [2003b]. Within TREC, the commonly used relevance type is topical relevance, which is associated with the “aboutness” of given documents.

In the remainder of this section, we will introduce these components in detail by introducing an example test collection, the TRECVID 2006 collection that is used for the evaluation of video retrieval methodologies. An introduction to TRECVID is given by Smeaton et al. [2006]. Note that we will limit our description on issues related to our work, hence ignoring other parts of TRECVID, such as different search tasks and other data collections.

TRECVID 2006 Test Corpus

The TRECVID 2006 corpus consists of approx. 160 hours of television news video in English, Arabic and Chinese language which were recorded in late 2005. The dataset

²⁻⁶http://ir.dcs.gla.ac.uk/test_collections/, last time accessed on: 14 April 2010

²⁻⁷<http://www.imageclef.org/datasets/datasets>, last time accessed on: 14 April 2010

²⁻⁸<http://trec.nist.gov/data/blog.html>, last time accessed on: 14 April 2010

also includes the output of an automatic speech recognition system, the output of a machine translation system (Arabic and Chinese to English) and the master shot reference. Each shot is considered as a separate document and is represented by text from the speech transcript. In the collection, we have 79484 shots and 15.89 terms on average per shot, with 31583 shots without annotation. For each shot, representative key frames are provided. Further, the collection contains a set of predefined search topics and assessment lists for each topic. We elaborate this further in the next two sections.

Search Topics

According to Voorhees [2005], search topics have two purposes. First of all, they describe an information need. Hence, they are input to retrieval systems. Second, they guide human assessors when judging relevance of the output. The TREC Vid 2006 collection contains a set of 24 topics. Within TREC Vid, search topics consist of “multimedia statements”, consisting of a title, brief textual descriptions of the information need, a set of example key frames and example video shots. The topics express the need for video concerning concepts such as people, events, locations, things and combination of these concepts. An example title, taken from the 2006 corpus is “Find shots of Saddam Hussein with at least one other person’s face at least partially visible.” As can be seen from this example, the unit of retrieval within TREC Vid is a video shot.

Assessment List Generation

Aiming to evaluate the output of retrieval systems using these search queries, statements about the relevance of the retrieved documents are required. As argued above, we focus in this work exclusively on topical relevance, which is the main relevance type considered within TREC. According to Voorhees [2001], assessment lists used in TREC are typically binary; a document is either relevant or not relevant to the given topic. A simple approach of creating assessment lists is to manually assess the documents of test collections. A problem is, however, that relevance is relative. Cuadra [1967] have shown that even though when assessors are asked to assess relevance of documents to a given search task, they will most probably judge the relevance of these documents differently. Another problem is that, considering the large human effort involved, this approach is very expensive and therefore not suitable for large-scale collections. Spärck-Jones and van Rijsbergen [1965] argue for the creation of assessment lists using subsets of the actual collection. Assuming that the highest ranked documents of multiple independent retrieval runs will contain a large number of relevant documents, they propose to merge these results in a “pool” of documents. Assessors are then asked to judge relevance

of these documents. This approach, referred to as *pooling*, is the primary assessment method within TREC and TRECVID. The TRECVID collection contains assessment lists of 60 to 775 relevant documents for every search topic. [Sanderson and Joho \[2004\]](#) evaluate various other approaches which can compete with the pooling approach. None of the introduced assessment approaches, however, result in complete lists containing all relevant documents of the collection.

Evaluation Measures

A basic assumption to identify better systems is that better systems provide better ranked result lists. A better ranked list satisfies the user overall. In the last thirty years, a large variety of different evaluation measures have been developed to evaluate the retrieval system's ability to correctly associate documents to a given query. A detailed survey is given by [\[van Rijsbergen, 1979, chap. 7\]](#). The measures introduced in the *Cranfield II experiments* are recall and precision. They are nowadays the de facto main evaluation metrics of IR systems.

$$\text{Precision} = \frac{\# \text{ relevant documents retrieved}}{\# \text{ retrieved documents}}$$

Precision is a measure of the proportion of retrieved relevant documents. It is important in information search. Considering that users often interact with few results only, the top results in a retrieved lists are the most important ones. An alternative to evaluate these results is to measure the precision of the top- N results, $P@N$. $P@N$ is the ratio between the number of relevant documents in the first N retrieved documents and N . The $P@N$ value focuses on the quality of the top results, with a lower consideration on the quality of the recall of the system.

$$\text{Recall} = \frac{\# \text{ relevant documents retrieved}}{\# \text{ relevant documents in the collection}}$$

The recall measures the proportion of relevant documents that are retrieved in response to a given query. A high recall is important especially in copyright detection tasks.

Both precision and recall values are single-value metrics that consider the full list or retrieved documents. Since most retrieval systems, however, return a ranked list of documents, evaluation parameters should allow to measure the effectiveness of this ranking. One approach to combine these metrics is to plot precision versus recall in a curve.

Another popular summary measure of ranked retrieval runs is the “average precision” (AP):

$$AP = \frac{1}{\# \text{ relevant}} \sum_{k=1}^{\# \text{ relevant}} (\text{Precision at rank of } k^{\text{th}} \text{ relevant document})$$

The most popular single-value metric in the IR community is the arithmetic mean of average precision (AP) over all queries, the “mean average precision” (MAP). MAP evaluates precision at all recall levels. A condition for this measure is, however, that every search query is considered equal.

Note that the definition of “document” depends on the actual unit of retrieval. As mentioned before, the unit of retrieval in TRECVideo is a video shot. We therefore refer to video shots as documents.

Discussion

In this section, we introduced system-centred evaluation. Given a test collection, pre-defined search topics and assessment lists, the output of retrieval systems is evaluated using well-established evaluation measures. It is the most common evaluation scheme in the IR community and is mainly applied for batch evaluation, i.e. to fine tune system parameters or to evaluate algorithms that do not require user input. Moreover, the introduced data collection, search topics and assessment lists are used in interactive IR experiments. Indeed, interactive IR experiments form part of the TREC campaigns, e.g. in the “Interactive Search Task” within TRECVideo. Even though the evaluation scheme has been broadly accepted as evaluation standard, its design has various flaws. A thorough discussion on its drawbacks is given in [Spärck-Jones, 1995, 2000; Blair, 2002; Hersh et al., 2000]. One of the main critique points is the evaluation scheme’s controlled experimental design, which can affect the user’s behaviour while using the retrieval system during a user study. Robertson et al. [1997] argues that even though TREC-like data collections can be used to evaluate interactive information retrieval approaches, problems of “reconciling the requirements of the laboratory context with the concerns of interactive retrieval are still largely unresolved.” Parts of these problems can be addressed by evaluating systems using a user-centred evaluation scheme. We introduce this approach in the next section.

Facing these critique points, we agree with Voorhees [2006] that “no one pretends that test collections are perfect or all-wise. Indeed, it has been said that test collections are terrible for IR research except that they’re better than current alternatives”.

2.3.3 User-centred Evaluation

Robertson and Hancock-Beaulieu [1992] argue that system-centred evaluation is not

suitable for interactive IR systems, since the controlled evaluation environment ignores essential factors of human-computer interactivity. First of all, they criticise the idea of pre-defined relevance lists. They argue that relevance should consider user's personal information need. Moreover, they point out that user's interactions should be included into the evaluation process. Another evaluation paradigm that addresses above problems, inspired by both Human-Computer Interaction and Psychology, is user-centred evaluation. [Borlund \[2003a\]](#) refers to this model as IIR (interactive information retrieval) evaluation. Within this scheme, the user's perception and behaviour is the centre of the evaluation rather than system performances that can be measured by precision and recall. [Bailey et al. \[2005\]](#) argue that designing, running and analysing user studies is substantially more complex than simply comparing empirical evaluation measures as common in system-centred evaluation. Nevertheless, real user studies are an essential part in the evaluation of IR systems, since only then, the impact of research methodologies can truly be assessed. The main condition for performing user studies in the context of information access is to carry out user-centred evaluation in an unbiased and appropriate manner. In this section, we first discuss experimental settings of user-centred evaluation. Then, we introduce different evaluation measures, namely usage log file analysis and questionnaires.

Experimental Evaluation Framework

Aiming to address the main critique points toward the disadvantages of interactivity in the system-centred evaluation scheme, [Borlund \[2003a\]](#) introduces a framework for the evaluation of interactive information retrieval systems. Her framework can be seen as the de-facto standard evaluation framework for interactive IR systems. She argues that interactive IR systems should be evaluated under realistic conditions, i.e. the evaluation procedure should model actual information seeking tasks. Therefore, she suggests to recruit potential users as test subjects of the IR systems in question. Further, she propagates the idea of employing a "simulated search task" situation, where a search topic is set into context. Simulated search tasks are "cover stories" that describe a situation where a certain information need requires the use of an IR system.

According to [Campbell and Stanley \[1963\]](#), a well-known problem in evaluations involving humans is the humans' learning aptitude. Humans learn how to handle a system the longer they use it. Hence, results of subsequent experiments most likely will be better than the results of early experiments. Besides, users might be familiar with a specific topic and will return better results than unexperienced users without any background knowledge. A well-established evaluation pattern to address this problem is the

2.3. Evaluation Methodologies

so-called Latin-Square evaluation design where user and topic are treated as blocking factors. Imagine, for example an experimental methodology where the effectiveness of two interactive IR systems (V_1) and (V_2) shall be measured. Assuming an equal number of search topics, each user would perform half of these topics. The approach allows the estimation of effectiveness of one system, free and clear of searcher and topic. However, it does *not* solve cross-site comparisons problems. Table 2.1 shows a 2×2 latin square design.

TABLE 2.1: 2×2 latin square design

	T_1	T_2
S_1	V_1	V_2
S_2	V_2	V_1

It has to be interpreted in the following way:

- Searcher S_1 uses System V_1 for Topic T_1 and System V_2 for Topic T_2 .
- Searcher S_2 uses System V_2 for Topic T_1 and System V_1 for Topic T_2 .

Asking users to perform such simulated search tasks, their interactions can be captured in usage log files. Further, their valuable feedback can be gathered in questionnaires. Both information sources will be introduced in the remainder of this section.

Usage Log File Analysis

Usage log files, also referred to as transaction log files, contain a recording of the communications between users and the system they are interacting with. [Rice and Borgman \[1983\]](#) define them as an automatic data collection method that captures the type, content and time of transactions. [Peters \[1993\]](#) presents logs as electronically recorded interactions between information retrieval systems and the persons using these systems. They are a good method to unobtrusively collect a robust set of data on a sizable number of system users. Information about user and system interaction can be gathered *without* interrupting the information seeking process. The row caused by the release of the American Online (AOL) query logs in 2006 illustrates the rich content of respective log files. Hence, it is a popular technique among researchers to evaluate retrieval systems (e.g. [[Croft et al., 1995](#); [Jansen et al., 2000b](#); [Jones et al., 1998](#); [Wang et al., 2003](#)]). Web search engine companies use log files to improve their retrieval systems. [Jansen \[2006\]](#) addresses the transaction log as a research methodology.

To obtain information from log files, the data needs to be analysed. In literature, this process is commonly referred to as *transaction log analysis* (TLA). This approach is based on the systematic affirmation of hypotheses in comparing and sampling data [Glaser and Strauss, 1967]. Considering arising privacy issues, Adar [2007] argues to anonymise log files. As pointed out by Drott [1998], a log file analysis can address various research questions and interaction issues. However, it usually focuses either on information structure, measurement of user interaction or system performance. Peters [1993] defines the analysis as the study of recorded interactions between information retrieval systems and its users. The aim of a TLA is to understand the interactions between users, content and the retrieval system. Or, dependent on the research question, the interaction between two of these elements. Possible achievements can be the confirmation of a research hypothesis, indices on the lack of the applied interface features or a better understanding of the users' searching behaviour.

Questionnaires

Bulmer [2004] defines questionnaires as a "structured research instrument which is used to collect social research data in a face-to-face interview, self-completion survey, telephone interview or Web survey. It consists of a series of questions set out in a schedule, which may be a form, on an interview schedule on paper, or on a Web page." Both the interview and self-completion survey (electronic or via pen-and-paper) questionnaire modes are commonly used to gather user opinion about interactive information retrieval experiments. Various question types are most commonly used:

- **Open questions:** They are useful to find out more about the reasons, *why* users behave the way they do and provide the chance to give free comments on aspects of the system. In IR experiments, they are used e.g. to gather the users' opinion about a specific feature of the system. Furthermore, they are used to identify positive and negative features from the users' point of view.
- **Closed questions:** Users can respond to a given set of responses. They can be in the form of a statement such as "I was satisfied with the results of my search". The Five-Point *Likert Scale* technique is taken for quantifying the expression of agreement or disagreement of a user. It presents a set of attitudes. For measuring the level of agreement, a numerical value from one to five is used. The value can be measured in calculating the average of all received responses. The other type of structured question, the *semantic differentials* provide a set of bipolar adjectives with a five-step rating scale between them. The adjectives can express one's attitudes.

As each type of question has its advantages and disadvantages, a combination of both question types is commonly used.

According to [Czerwinski et al. \[2001\]](#), users rarely evaluate computer systems poorly. They tend to inflate their ratings during usability evaluations. Such effects occur due to using a particular questionnaire technique. Studies have revealed a complex relationship between the questionnaire type, question content and users' responses. [Tourangeau et al. \[2000\]](#) pose that users are more willing to report sensitive information in self-completion surveys than in interviews. A popular explanation for this phenomenon is social desirability, or, as described by [Richman et al. \[1999\]](#): “the tendency by respondents [...] to answer questions in a more socially desirable direction that they would under other conditions or modes of administration”. [Kelly et al. \[2008\]](#) investigated the relationship between users' responses and the questionnaire type during an interactive IR experiment. They concluded that for open questions, the pen-and-paper method is the most efficient mode to gather information. Besides, their research showed that users' quantitative evaluation were significantly higher using an electronic questionnaire than in a face-to-face interview. They suggest to use face-to-face interviews for closed questions and either electronic or pen-and-paper techniques for open questions.

Discussion

In this section, we introduced the user-centred evaluation scheme where user studies are conducted to perform the effectiveness of interactive IR systems. As we have shown, usage log files and questionnaires are used as evaluation measure. The introduced experimental methodology consisting of pre-defined search topics and (varieties) of the Latin-Square evaluation design are the de-facto evaluation standard for interactive experiments within TREC. We therefore apply the user-based evaluation scheme in this work.

[Christel \[2007b\]](#) criticise that in the interactive video retrieval domain, most research approaches focus on short-term retrieval as advocated within the TRECVID evaluation campaign, hence ignoring more realistic video retrieval scenarios. Considering the broader focus of realistic video search, they argue for “Multi-dimensional In-depth Long-Term Case-studies (MILC)”, as advertised by [Shneiderman and Plaisant \[2006\]](#). Multi-dimensional stands for different evaluation measures, including interviews, surveys and logging user interaction to measure system performance. In-depth analysis aims to include the researcher into the study process, e.g. by assisting subjects. Long-term refers to longitudinal studies, where users interact with a system over multiple sessions. Case studies aim to set the evaluation into realistic scenarios, e.g. in the users'

natural environment. Long-term user studies are very common in the HCI community, but only have recently drawn the attention of the IR community. Examples include Kelly [2004], who studies users' online behaviour over a time period of fourteen weeks. Every week, users were asked to fill in a questionnaire, where they had to evaluate documents they interacted with. Further, interaction logs were used to evaluate their interactions. Similarly, Lee et al. [2006] evaluates a news video retrieval system over a period of one month without any supervision. Participants of their study were asked to use the system to satisfy their personal information needs and to keep a diary about their experiences with the system during the study. Both log files and questionnaires were used to evaluate their hypotheses. Liu and Belkin [2010] conduct a two weeks experiment where their participants were asked to perform search tasks under the authors' supervision. Thus, they evaluate long-term personalisation techniques under a controlled lab conditions. Log files and questionnaires are employed as evaluation measure. Within this thesis, we will study the usability of implicit relevance feedback over multiple sessions. Thus, we will base our evaluation on the introduced work.

2.3.4 Simulation-Based Evaluation

Within the last twenty years, automated evaluation of non-interactive IR systems and approaches as introduced in Section 2.3.2 has been well-established in the IR community. Considering, however, that many real-life IR systems adapt their retrieval results based on the users' feedback and system usage, such automated evaluation scheme cannot easily be applied in interactive information retrieval [White et al., 2005; Belkin, 2008]. In the previous section, we introduced conditions for a user-centred evaluation scheme where the user is included into the evaluation of interactive information retrieval and recommender systems. User-centred evaluation schemes are very helpful in getting valuable data on the behaviour of interactive search systems, however, they are expensive in terms of time and money, and the repeatability of such experiments is questionable. It is almost impossible to test all the variables involved in an interaction and hence compromises are required on many aspects of testing. Furthermore, such a methodology is inadequate in benchmarking various underlying adaptive retrieval algorithms. An alternative, well-established way of evaluating such systems is the use of simulations, i.e. automated evaluation runs that consider the user and their interaction with the retrieval systems. Surveys on the simulation-based evaluation of computer systems include Ivory and Hearst [2001]; Zeigler [1984]; Hartmann [2009]. In this section, we provide an overview on most recent approaches that are most relevant to our own work, namely simulation of relevance feedback and interface evaluation.

Simulation of Relevance Feedback

White et al. [2005] proposed to evaluate the performance of various implicit indicators in the text retrieval domain by employing a simulation-based evaluation scheme. Within their simulation, they model a user who interacts with initial search results, following different information seeking strategies: (1) the user shows only interest in relevant *or* non-relevant documents, (2) the user views *all* relevant or non-relevant documents, (3) the user shows interest in relevant or non-relevant documents to a different degree. Since they base their experiment on a TREC collection, they exploit the relevance assessments of the collection to identify relevant and non-relevant documents. A similar study is introduced in White [2006], where log files of a preceding user study are used to determine the proportion of different information seeking strategies.

Keskustalo et al. [2008] evaluated the effectiveness of simulated relevance feedback by modelling a simplified user interaction scenario. The interaction scenario can be split into two phases: Firstly, an initial search query is triggered and a simulated user provides feedback on retrieved results. Secondly, they expand the initial search query with terms extracted from the feedback document, trigger a new retrieval and evaluate the returned document list. They evaluate different stereotype user types that provide different grades of feedback quality: An impatient user, a moderate user and a patient user. A follow-up experiment is conducted by Järvelin [2009] who evaluate whether relevance feedback returns better results than pseudo relevance feedback approaches. For both studies, TREC collection with pre-assessed relevance assessment lists are used to evaluate their hypotheses.

Joho et al. [2009] exploit the log files of a preceding user study with the purpose of evaluating a number of IR techniques applied to collaborative search. Their main research question was whether relevance feedback can be effectively employed in a collaborative scenario. They first create a pool of search queries that had been formulated by various users while performing the search tasks. Exploiting this pool, they simulate users collaboratively performing a search task over up to ten iterations. They evaluate eight different search strategies where individual retrieval results are influenced to various extends based on the simulated partner's previous retrieval results. They employ a TREC collection for their study, thus using the assessment lists to evaluate the approaches. Foley and Smeaton [2008] follow a similar approach in their simulation of a synchronous collaborative information retrieval environment. They exploit log files of independent search sessions by synchronising the start time of different search sessions and analyse how different events in the log files could have influenced the retrieval process.

A different approach is introduced by [Shen et al. \[2005\]](#), who exploit clickthrough data of a small user study to evaluate several context-sensitive retrieval algorithms. Identifying preceding search queries within a search session, they compare the retrieval results of these search queries with various context-sensitive search queries. Thus, they simulate users by repeating their query history. [Hopfgartner et al. \[2008b\]](#) follow a similar idea by exploiting search query histories of a user study to evaluate different techniques to improve video retrieval. Both studies are conducted with TREC collections and thus, standard evaluation measures are used to evaluate their methodologies.

[Keskustalo et al. \[2009\]](#) criticise that even though real users rely on a series of very short search queries rather than complicated longer queries, Cranfield style IR experiments mostly evaluate approaches consisting of long and complicated search queries. They evaluate the performance of a sequence of short search queries by employing a simulation-based evaluation scheme. Given a list of TREC search topics, users were asked to name potential search terms that could be used to retrieve the corresponding topic. Further, they were asked to form search queries consisting of one, two, three or more of these terms. They then simulate users searching for the corresponding search task of the TREC collection by triggering sequences of various query combinations. Results of each iteration are then evaluated using standard evaluation measures.

[Dix et al. \[1993\]](#) argued that user interactions consist of a series of low-level events such as key presses or system reactions. They reason that any task performed by users can be represented by hierarchically combining these low-level events and that this interaction representation can be described as finite state machines. State changes are triggered with a certain probability by certain events such as other user actions or system responses. State transitions can be identified by applying machine learning techniques. An example is given by [Bezold \[2009\]](#), who defines user interactions with an adaptive interactive system in a probabilistic deterministic finite-state automaton where user actions are represented as states within the automaton. The author argues to determine the probabilities by performing a log file analysis. [Wilson et al. \[2009\]](#) argue on the same lines and propose to measure strengths and weaknesses of search interface designs by analysing possible information seeking patterns. Even though they do not conduct a simulation-based evaluation, their evaluation framework can be seen as a guideline on how to perform user simulations.

Simulation of Browsing Patterns

Above simulation methodologies are defined to evaluate underlying IR approaches that depend on users providing relevance feedback. Another application of simulation-based

evaluation focuses on studying different interface designs.

[Chi et al. \[2001\]](#) argue that web surfing behaviour is the users' mean to express their information needs. They analyse web surfing patterns of visitors of a Web site and mimic this surfing behaviour to satisfy this need. Given a specific information need, they compute the probability of a user clicking on a web link. The probability is determined by analysing the target document's content similarity to the user's information need using TF.IDF. Since they do not use a standard test collection, their user model is evaluated by human assessors who rate the quality of the selected documents.

[Smucker and Allan \[2006\]](#) simulate users' browsing behaviour when interacting with a hypothetical retrieval interface. They model two different patterns: A greedy pattern and a breadth-like pattern. The greedy pattern simulates users clicking on every relevant search result to retrieve similar documents of that document. The breadth-like pattern simulates users inspecting the results' snippets in the list first and retrieving similar documents for the most relevant documents only. Technically, they model this pattern by adding all relevant retrieval results to a first-in-first-out-queue until precision at N , where N is the rank of the current document, is below a predefined threshold and then retrieve similar documents for all documents in the queue. [Lin and Smucker \[2008\]](#) apply user simulations to measure interface utility following the same methodology of mimicking different browsing patterns. In both approaches, standard evaluation measures apply. A similar methodology was introduced by [Warner et al. \[2006\]](#) who simulate users interacting with a basic graphical user interface showing a ranked list of web links. The simulated user interacts with this ranked list, i.e. they simulate clicks on the web links. They do not distinguish between relevant or non-relevant documents, the decision to simulate a click is based on randomised parameters.

In the image browsing domain, [Leelanupab et al. \[2009a\]](#) simulate users browsing an image collection. Analysing the log files of a preceding user study, they identify possible user interactions such as triggering a new search query, browsing images or starting new search sessions. In a first step of their simulation strategy, they simulate a user triggering a search query that has been identified from the log files of mentioned user study. Repeating the search session of this search query, their simulated user interacts with a limited number of retrieved images, i.e. they browse through images similar to the image. Since they did not use a pre-assessed test collection, they evaluate the outcome of their simulation using statistical significance tests.

[Hopfgartner et al. \[2010b\]](#) perform a simulation-based evaluation scheme to evaluate the performance bounds of an aspect-based video retrieval system that allows users to trigger parallel retrieval sessions and to re-arrange results between these sessions. Their simulation starts with an initial search query. Arguing that an initial search query

has a high probability of being general, thus containing a diverse set of documents, they first cluster the result list to obtain coherent semantically related aspects. They assume that the top k clusters form the k facets of a user's information need and use them to create more specific queries. These queries are then used to automatically propose new sets of results. Finally, their iterative clustering process is used to identify new aspects and refine the queries and consequently the retrieved results. They evaluate their approach using the TRECVID data corpus, thus using precision to evaluate different querying approaches.

Discussion

In this section, we surveyed simulation-based evaluation, an evaluation paradigm that is based on simulating user interactions to evaluate interactive information retrieval approaches. We focused on two main applications: Simulation of relevance feedback, browsing and interface evaluation.

As we have shown, most simulation schemes rely on pre-defined interaction patterns, often backed by statistical click analyses. Stereotype users are mimicked, e.g. by analysing how often and under which conditions actions are performed by real users. Most simulations are rather generic and based on heuristic user interactions. Due to these limitations, we agree with [White et al. \[2005\]](#) that user simulations should only be seen as a pre-implementation method which will give further opportunity to develop appropriate systems and subsequent user-centred evaluations.

Within this work, we therefore apply simulation-based evaluation schemes that are based on the above introduced methodologies. First of all, we agree with [Dix et al. \[1993\]](#) that users' interactions with retrieval interfaces can be seen as low-level events. Following their argumentation, we argue that providing relevance feedback can be seen as low-level events within a user's information seeking task. Thus, we will model basic user interactions to simulate the role of relevance feedback in the video retrieval domain. Above survey has shown that two user modelling approaches dominate the research field. Either, simplistic behaviour patterns are defined based, e.g. by analysing the user interfaces or behaviour patterns are determined by analysing log files. Within this work, we will rely on both approaches to study various hypotheses. Moreover, following most of the above introduced approaches, we aim to employ standard evaluation measures for our evaluation.

2.3.5 Summary

In this section, we surveyed standard evaluation methodologies within the IR domain. As pointed out, evaluation of IR systems can be segmented into three main methodologies. System-centred evaluation focuses on evaluating system parameters and approaches using standard evaluation measures. We introduced the controlled evaluation paradigm referred to as Cranfield Evaluation Paradigm, consisting of closed test collections, various search topics and corresponding assessment lists. Further, we introduced the classical evaluation measures. User-centred evaluation concentrates on measuring user behaviour and satisfaction. We first introduced Borlund's IIR evaluation model and discussed issues for the design of unbiased user studies. Differing from system-centred evaluation, user studies can be studied by analysing usage log files and questionnaires. The last evaluation methodology aims to simulate user interaction. It is employed whenever a real user study is not suitable, e.g. in large-scale evaluation of system parameters that require user interaction.

We discussed strengths and weaknesses of introduced approaches. Summarising, we conclude that evaluation in IR is, due to the TREC initiative, very well-defined. We therefore apply these evaluation schemes to evaluate the research questions and hypotheses within this thesis.

– *We are the heirs of our own actions.*

Buddha, 563–488 B.C.

3

The Role of Implicit Relevance Feedback

As discussed in the previous chapter, implicit relevance feedback is a well-studied approach for adapting search results in the text retrieval domain. This chapter focuses on exploiting implicit relevance feedback for short term user profiling in the video domain, i.e. for recommending relevant videos within a search session. Section 3.1 introduces the research problem. Implicit indicators of relevance are identified in Section 3.2. In Section 3.3, we model different user actions by analysing representative video retrieval interfaces. Section 3.4 introduces a simulation-based evaluation scheme where users' interactions are simulated. Results of this simulation are presented in Section 3.5 and discussed in Section 3.6.

3.1 Introduction

White [2004] has shown that *implicit relevance feedback* can successfully be employed to support text retrieval tasks. By mining implicit user interaction data, it is possible to infer user intentions and retrieve more relevant information. Traditional issues of implicit feedback can also be addressed in video retrieval since digital video libraries facilitate more interaction and are hence suitable for implicit feedback.

Graphical user interfaces of both textual and multimedia domains are designed to assist users in their information seeking task. Considering that each interface feature is designed to allow users to either retrieve or explore document collections, we hypothesise that the users' interactions with these features can be exploited as implicit relevance

feedback (Hypothesis H_1). Dix et al. [1993] argue that user interactions in interactive systems can be represented as a series of low-level events, e.g. key presses or mouse clicks. Any action that users perform during their information seeking activity consist of a series of these events. A first necessary step to support Hypothesis H_1 , that low-level feedback events of video retrieval interfaces can be exploited as implicit indicator of relevance is to identify these events. The first contribution of this chapter is therefore the identification of possible low-level events. In Section 3.2, we analyse representative state-of-the-art video retrieval interfaces to identify these events.

As discussed in Section 2.1.5, the specific nature of video data requires rather complex graphical user interfaces. Consequently, a large variety of different interface designs exist and thus, the way users interact with these interfaces and provide implicit relevance feedback differs significantly from their textual counterparts. Different interface designs result in different user interactions with these interfaces, triggering different implicit indicators of relevance which could be used to infer relevance. Extending the initial hypothesis, we therefore hypothesise that the interpretation and the importance of these events depend on the interface context (Hypothesis H_2). Section 3.3 models various user action sequences that illustrate the effect that different interface designs have on a user's search behaviour.

An interesting research challenge is how to evaluate both hypotheses. As discussed in Section 2.3.4, a common approach for studying the users' behaviour of interacting with a computer system is to perform a user study, to monitor the users' interactions and to analyse the resulting log files. Such an analysis shall help to identify good implicit indicators of relevance, as it can help to answer basic questions: What did the user do to find the information he/she wanted? Can the user behaviour be used to improve retrieval results? In order to get an adequate impression of the users' behaviour when interacting with a video retrieval system, two main criteria can be stressed out. A large quantity of different users interacting with the system is necessary to draw generalisable conclusions from this study, i.e. by analysing user log files. In addition, non-expert users should be interacting with the systems, as they will interact in a more intuitive way than expert users. However, such a methodology is inadequate for the evaluation of interactive retrieval systems. Most interactive video retrieval systems are evaluated in laboratory-based user experiments. There are many issues with such evaluation methodologies such as the lack of repeatability. In addition, to achieve a robust measurement, we need a large user population, which is very expensive. Furthermore, it is difficult to benchmark different parameter combinations of features for effectiveness using user-centred evaluations. An alternative way of evaluating such user feedback is the use of simulated interactions. Analysis of the research efforts that have been sur-

veyed in Section 2.3.4 lead to the conclusion that even though simulation-based studies should be confirmed by user studies, they can be a cheap and repeatable methodology to fine tune video retrieval systems. Hence, user simulation is a promising approach for further study of adaptive video retrieval, at least as a preliminary step. The second contribution of this chapter is, therefore, a simulation-based evaluation scheme aiming to support both hypotheses. Section 3.4 presents a scheme that can also be used as a preliminary methodology for the study of implicit relevance feedback. Results are introduced in Section 3.5 and discussed in Section 3.6.

In summary, this chapter aims at evaluating the following hypotheses:

H_1 : Implicit relevance feedback can be employed to support interactive video retrieval.

H_2 : The interpretation and importance of the implicit indicators of relevance depend on the interface context.

The research which is presented in this chapter has been published in [Hopfgartner et al., 2007; Hopfgartner and Jose, 2007].

3.2 Low-Level Feedback Events of Video Retrieval Interfaces

Dix et al. [1993] define user interactions with interactive systems as a series of low-level events. These events are the most basic interactions that users can perform during their interaction. The first requisite to study the role of implicit indicators of relevance is to identify these low-level events in the video domain. In Section 2.1.5, we surveyed representative graphical user interfaces. As mentioned, a larger survey is presented in [Schöffmann et al., 2010]. The surveyed interfaces provide various low-level feedback events that users can trigger while interacting with given documents. Any action that users perform during their information seeking activity, further referred to as their search session, consists of a *series* of these events. As stated with Hypothesis H_1 , we assume that these events can be used for implicit relevance feedback. The following six events have the potentials to be exploited as implicit indicators of relevance in the video domain:

- *Previewing*: Hovering the mouse over a key frame. This can result in a tooltip showing neighbored key frames and additional text or in highlighting the query terms in the text associated with the key frame. This low-level event indicates further interest in a key frame as the user receives additional information about the result.

- *Clicking result*: Click, e.g. on a key frame, to trigger playback of a video shot or to perform further actions. This event indicates the users' interest in the video shot which is represented by the key frame.
- *Sliding*: Using the sliding bar to navigate through a video. This event indicates further interest in the video. Users appear to slide through a video when the initial shot is not exactly what they were searching for but when they believe that the rest of the video *might* contain other relevant shots. Hence, the initial shot might not be an exact match of the users' need but raises hope to find something of relevance in the same video.
- *Exploring*: Looking at metadata (date of broadcast, broadcasting station,...). By performing this event, users show a higher interest in the current shot, as they want to get additional information. This information can help them to judge about the relevance of the shot. A user for example might search for a specific sports event such as the football world cup final. In such cases, the direct correlation between broadcasting date and event date can help to identify relevant shots as such events usually appear in the news shortly after their happening.
- *Browsing*: Browsing through a video by clicking on its neighboured key frames. Similar to using the sliding bar to navigate through a video, this feedback indicates users' interest in this shot. Unlike using the sliding bar, browsing indicates that users suspect a relevant shot in the neighbourhood of the current shot.
- *Viewing*: Viewing a video. The playing duration of a video might indicate users' interest in the content of the video.

The main research challenge that arises when using these low-level events as implicit indicator of relevance is that it is not clear whether these events are positive or negative indicators for relevance. In the text retrieval context, [Claypool et al. \[2001\]](#) identified time spent on a web site as being a valid implicit indicator of relevance in the text domain. [Kelly \[2004\]](#), on the other hand, criticises the time factor as implicit indicator. She assumes that information-seeking behaviour is not influenced by contextual factors such as topic, task and collection. Thereupon, we state in Hypotheses H_2 that the interpretation and the importance of these indicators depend on the interface context. Viewing a video might, for instance, be essential in one interface, but of supportive nature only in another interface. In the next section, we highlight this challenge further by introducing different action sequences that highlight how users could interact with a given document using representative video retrieval interfaces.

3.3 User Action Sequences

As stated above, interactions are defined as a series of low-level events which are performed by users using the video retrieval systems. Bezold [2009] describes such event series as probabilistic finite-state automata. Considering that each low-level event combination within a user’s interaction sequence depends on the preceding event, we argue that user interactions can be simplified in a Markov Chain [Meyn and Tweedie, 1996]. Markov Chains consist of states and transitions between these states. A state change is triggered by a certain event with a certain probability. In this section, we introduce five Markov Chains that represent possible user action sequences consisting of low-level events when users interact with a given document using interactive video retrieval systems. Aiming for a preliminary evaluation of the role of these events, not all possible events provided by each interface are integrated in the sequences. Hence, a scenario covers some *possible* user interaction, not necessarily a user interaction including *all* features the interface provides.

3.3.1 User Action Sequence S_1 (Preview-Click’n’View)

Sequence S_1 combines three different low-level events, encompassing all interfaces that provide the minimal functionalities of previewing, clicking on a key frame in the result set and viewing the video shot. Due to these functionalities, we refer to this sequence as “Preview-Click’n’View”. Example interfaces that allow this event combination have been presented in Christel and Conescu [2005]; Hopfgartner et al. [2007]. Possible low-level event combinations are visualised in Figure 3.1.

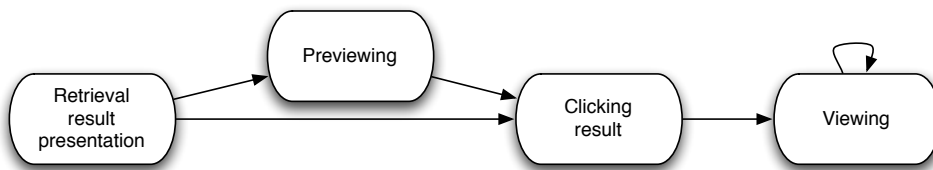


FIGURE 3.1: Possible event combinations on a given document in Sequence S_1

Given a displayed document, denoted “Retrieval result presentation” in above figure, this sequence models users (i) hovering the mouse over listed key frames to get some additional information of the shot, e.g. in a tooltip (previewing). Further, the users may (ii) click on the key frame (clicking result) to (iii) start playing a video (viewing).

These actions result in the use of the following implicit indicators of relevance:

- i. Previewing
- ii. Clicking result to trigger video playback
- iii. Viewing

3.3.2 User Action Sequence S_2 (Click'n'Browse)

Sequence S_2 , referred to as “Click’n’Browse”, combines two low-level events that can be given when interacting with a document: clicking a result on a result list to display the video and its key frames, followed by browsing these key frames. An example interface supporting this sequence is introduced by Heesch et al. [2004]. In this interface, information is presented on different panels. Retrieval results are represented by key frames. Clicking on one key frame in a result panel will set focus on that key frame and update all other panels. One panel contains the neighboured key frames in a fish eye presentation. In this panel, a user can browse through the results. Possible low-level event combinations are visualised in Figure 3.2.

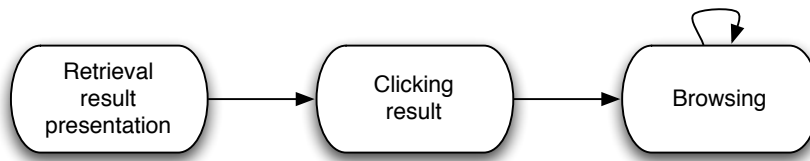


FIGURE 3.2: Possible event combinations on a given document in Sequence S_2

In this sequence, users can (i) click on a key frame in the result list (clicking result) and (ii) browse through its presented neighboured frames (browsing).

These actions result in the use of the following implicit indicators of relevance:

- i. Clicking result to update panels
- ii. Browsing through neighboured key frames

3.3.3 User Action Sequence S_3 (Click-View-Explore’n’Slide)

The third sequence S_3 covers an event combination which can be achieved when interacting with a document using the text-only video retrieval system provided by Browne et al. [2003]. Their web interface ranks retrieved results in a list of relevant video programmes. Each row displays the most relevant key frame, surrounded by its two

neighbourhood key frames. Below the shots, the text associated with the result is presented. The query terms which are associated with the key frame are highlighted when the user moves the mouse over the key frame. When clicking on a key frame, the represented video shot can be played. Different from S_1 and S_2 , this sequence considers two additional low-level events: highlighting metadata (exploring) and using a sliding bar (sliding). We refer to this scenario as “Click-View-Explore’n’Slide”. Possible low-level event combinations are visualised in Figure 3.3.

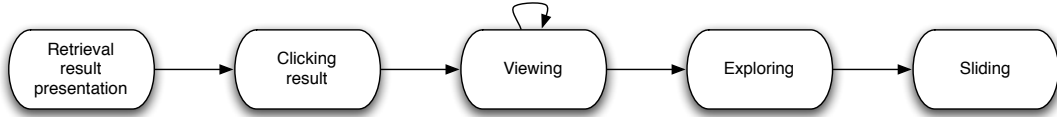


FIGURE 3.3: Possible event combinations on a given document in Sequence S_3

In this sequence, users can (i) click on a key frame (clicking result) to trigger (ii) video playback (viewing). They can (iii) highlight associated query terms (exploring) and (iv) navigate through the video using a sliding bar (sliding).

These actions result in the use of the following implicit indicators of relevance:

- i. Click result to trigger video playback
- ii. Viewing
- iii. Exploring
- iv. Sliding

3.3.4 User Action Sequence S_4 (Preview-Click-View’n’Browse)

This sequence models the users’ interaction on a given document using the system provided by Hopfgartner et al. [2007]. In their interface, retrieved video shots, represented by a key frame, are listed in a result panel. Hovering the mouse over a key frame will highlight a tooltip showing its neighbourhood key frames and the associated text (previewing). When clicking on a key frame, the corresponding video is played (viewing). The video which is currently played is surrounded by its neighbourhood key frames. Users can click on them and browse through the current video (browsing). We refer to this sequence as “Preview-Click-View’n’Browse”. Possible low-level event combinations are visualised in Figure 3.4.

In this sequence, users can (i) highlight additional information in moving the mouse over a retrieved key frame to get some additional information of the shot (neighbourhood

3.3. User Action Sequences

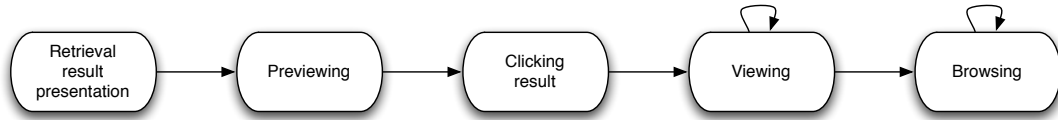


FIGURE 3.4: Possible event combinations on a given document in Sequence S_4

key frames and text from the speech recognition software) (previewing), (ii) click on a key frame of a result list (clicking result) and (iii) play a video (viewing). Also, they can (iv) browse through the video to find new results in the same video (browsing).

These actions result in the use of the following implicit relevance feedback:

- i. Previewing
- ii. Clicking result to trigger video
- iii. Viewing
- iv. Browsing

3.3.5 User Action Sequence S_5 (Click'n'More)

Sequence S_5 is the most complex of all introduced sequences. In contrast to the other sequences, it supports explicit relevance feedback. It is based on the retrieval interface by [Christel and Conescu \[2005\]](#). In this interface, retrieved results are represented by key frames and presented in a list. Clicking on one key frame, the user can choose to explicitly mark a shot as relevant (providing explicit relevance feedback), to play the video (viewing) or to display additional information (exploring). Further, it allows to browse displayed key frames (browsing) and slide through the video (sliding). We refer to this sequence as “Click’n’More”. Possible low-level event combinations are visualised in Figure 3.5.

In this sequence, users can (i) click on a key frame in the result list (clicking result) and (ii) play a video (viewing). They can also (iii) use the sliding bar (sliding). Users may (iv) browse through the video to find new results in the same video (browsing). Moreover, they can (v) show additional video information and sort results by date and broadcasting station (exploring). Besides, they can explicitly judge the relevance of a video shot (providing explicit relevance feedback).

These actions result in the use of the following implicit indicators of relevance:

- i. Clicking result to trigger video

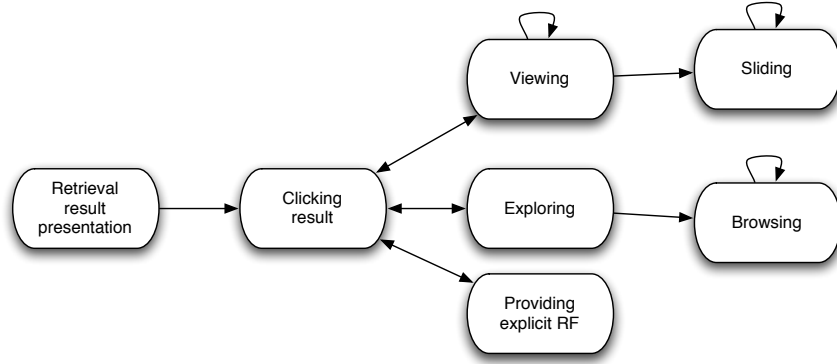


FIGURE 3.5: Possible event combinations on a given document in Sequence S_5

- ii. Viewing
- iii. Sliding
- iv. Browsing
- v. Exploring by listing metadata

3.3.6 Discussion

In this section, we modelled five user action sequences based on representative video retrieval interfaces that have been surveyed in [Schöffmann et al., 2010]. Each sequence models different user interaction scenarios where users trigger different events while interacting with a given document. The introduced user interaction scenarios illustrate that the design of graphical video retrieval interfaces directly influence user behaviour patterns. Even though users might follow the same aim, i.e. finding documents of interests, the interface design forces them to interact differently. Sequence S_2 (Click’n’Browse), for example, shows that users can interact with video results *without* viewing the actual video. In sequence S_3 (Click-View-Explore’n’Slide), however, viewing a video is essential while interacting with the results. Consequently, we argue that the interpretation and importance of the implicit indicators of relevance depend on the interface context (Hypothesis H_2). An interesting question is how these interaction patterns influence retrieval performance when the underlying low-level feedback events are exploited as implicit indicators of relevance. The next section aims to address this question by following a simulation-based evaluation scheme. By applying the above introduced scenarios, we simulate users providing implicit relevance feedback while performing a retrieval task over multiple iterations. Exploiting this simulated feedback,

we adapt retrieval results and measure the performance. Even though the simulation models a simplified retrieval scenario, we argue that the results can be seen as preliminary indications on how the low-level events can be exploited.

3.4 Simulation-based Evaluation

The scenarios which have been introduced in the previous section illustrate possible actions, consisting of low-level events, that users can perform while interacting with given documents using video retrieval interfaces. We argue that these low-level events can be used as implicit indicators of relevance, thus stating (Hypothesis H_1) that implicit relevance feedback can be employed to support interactive video retrieval. As the scenarios indicate, users interactions are directly influenced by the design of the interface, leading to Hypothesis H_2 that the interpretation and importance of the implicit indicators of relevance depend on the interface context. In order to evaluate both hypotheses, a thorough analysis of the role of the introduced implicit indicators is required, e.g. by following a user-centred evaluation scheme. This classical user study scheme would require multiple runs where users must use different user interfaces. Moreover, large-scale studies would be required to avoid external factors that could directly or indirectly influence the outcome of these studies. Since fulfilling these conditions is challenging, we presented in Section 2.3.4 the *simulation* of a classical information seeking process as an alternative methodology of evaluating such questions. Even though a simulation-based evaluation will not replace a real user study, it can nevertheless be used to draw preliminary conclusions. In this section, we therefore opt for the simulation-based evaluation. We will base our simulation on the TRECVID corpus. In Section 3.4.1, we first introduce how the classical TRECVID scenario can be modelled. The underlying feedback model is simple: Users' feedback and interactions with the system within a search session are used to identify relevant results. Once these results have been identified, additional search terms for query expansion are determined. The simulation is based on various parameters which are fine tuned in the Section 3.4.2. Section 3.4.3 concludes the section.

3.4.1 User Interaction Simulation

It is assumed that the identified implicit indicators of relevance can be used to identify documents which are relevant in a search task t . Further, we assume that the probability of a document being relevant is correlated to the attention, expressed by interactions, a user gives to this document within one search session.

Since we want to simulate a TRECVID like scenario, we rely on the 24 topics/queries associated with the TRECVID data set. The transcripts of the corpus, with stop words being removed and terms stemmed, are indexed using the retrieval engine Terrier [Ounis et al., 2005]. Search results are ranked using Okapi BM25.

In this simulation, we define a user session s as a set of Queries $Q_s = \{q_0, \dots, q_i\}$ which were input by the user u over $i \in I$ iterations. Further, we define a set of documents D_s that users interacted with within this search session and a set of low-level interaction events $e = \{\text{clicking result, previewing, exploring, viewing}\}$. Topical relevance of each document is defined by the relevance assessments for each TRECVID task. This allows us to simulate users interacting with a different number of relevant retrieved documents. For each iteration i , we execute a user query q_i and simulate various interaction events e of user u with the retrieved documents D_s as shown in Section 3.3. The number of relevant vs. non-relevant documents that the simulated user interacts with, determined by exploiting the assessment data, is a parameter within the simulation. Its effect is discussed later. In order to capture the users' interest in a document, we define a relevance weighting w_e for every single event e . The more interaction events are performed on a document, the higher the overall weighting of this document. Documents achieving the highest weighting are then exploited to generate a search query for the next iteration. Figure 3.6 depicts the simulation of one iteration $i \in \{1, 2, \dots, I\}$ for a task t .

In detail, for each user scenario $S_1 - S_5$, we simulate a search session as consisting of the following steps:

1. Execute a simulated query q_i .
2. Simulate user u interacting with the top n documents D_s of the returned result set.
3. Generate the query q_{i+1} for the next iteration.

In the remainder of this section, we discuss these steps in detail.

Execute a simulated query

In the first step, we execute a simulated query q_i , generated by the query generation module in a previous interaction simulation (Step 3). In case this is the first simulated iteration ($i = 0$), the simulated query is created from the description of Task t .

Simulate user

In the second step, we simulate user u interacting with the top n documents D_s of the returned result set. The relevant number of relevant vs. non-relevant documents that

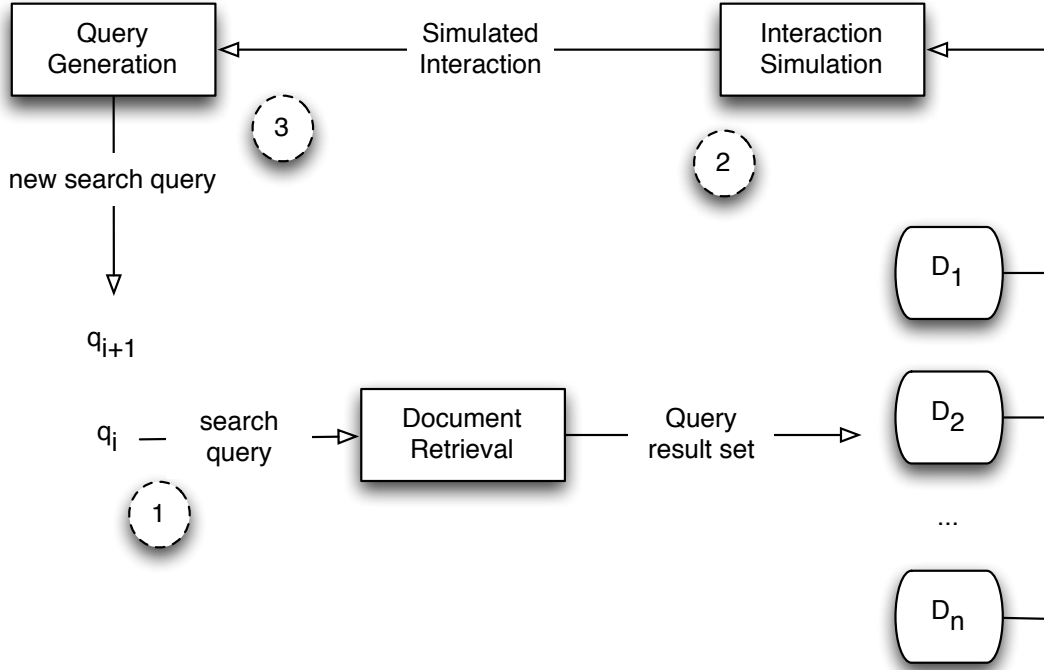


FIGURE 3.6: Steps and related components for a user interaction simulation

form n , i.e. the number of simulated interactions with relevant and non-relevant documents, can be set by exploiting the assessment data that form part of the TRECVID collection. The effect of different proportions of relevant vs. non-relevant documents is discussed in Section 3.4.2. The interaction consists of a series of events e as outlined in Section 3.3. In Markov Chains, transitions can be expressed by probabilities. Since we do not have any log files of previous studies that could be analysed in order to identify realistic transition probabilities between different actions in each event, we have to define various assumptions and conditions for this simulation. First of all, we assume that some event types, e.g. viewing a video for a longer period of time, appear more often in a user interaction work flow than others. A user can for instance view one video for five seconds and then view another video for ten seconds. We simplify this multiple appearance of this event by defining a minimum video viewing event. This could be for instance a user viewing a video for five seconds. If a video is viewed for a longer time period, a multiple instance of the minimum event is simulated. In the simulation, the viewing duration is limited to 0–10 cycles. Secondly, each low-level event will appear *randomly* in the simulation. This means that in each scenario, every transition between different states has a probability of 50%. Even though this is not a realistic model of a user, it should allow a preliminary analysis of the role of implicit indicators of relevance

in the video domain.

Further, we define a relevance weighting $w_e(D_s)$ for every event interaction performed on document D_s to verify the reliability of these low-level events as implicit indicators of relevance as follows:

$$w_e(D_s) = \begin{cases} n \in [0, 1) & \text{iff feedback on } D_s = \text{implicit} \\ 1.0 & \text{iff feedback on } D_s = \text{explicit} \end{cases}$$

Since explicit relevance feedback is the strongest indicator of relevance, any document which has been explicitly marked as relevant should have the highest weighting possible. We therefore define explicit feedback as the strongest relevance weighting with a maximum weighting of 1.0 given to all terms of the current document. Consequently, the sum of the implicit event weightings needs to be normalised to guarantee that the user feedback weighting will be between 0.0 and 1.0.

The simulation of “browsing” or “sliding” does not increase the weighting of the terms of the corresponding shot. Instead, it has an influence on the list of documents which are used for query generation. A description of the query generation process is provided in the next step. Assuming that a user mainly browses forward in time and rarely backwards, we model this browsing behaviour by considering the 0–10 right neighboured shots to the query expansion list. This simulates a user browsing 0–10 times to the right neighboured shot. In the “sliding” simulation, we simulate a user jumping randomly 0–10 times through the video. We take 0–10 random shots belonging to the same video and add them to the query expansion list.

A through analysis of weighting schemes is required to identify the importance of the actions that have been introduced before. This is, however, out of scope of this chapter, since we aim to study the effect that these indicators of relevance can have within the retrieval process rather than identifying the “best” indicators. Therefore, we define static weighting schemes for the different scenarios $S_1 - S_5$ within this study. The assigned values vary between 0.0 and 1.0 within the different sequences in order to avoid overvaluing specific low-level events. Each term in the document that the simulated user interacted with will be assigned to this weighting. Table 3.1 provides an overview of these schemes.

A simulation including Sequence S_1 , for example, is as follows: We simulate *pre-viewing*, *clicking result* to trigger the video playback and *viewing* that video for three time cycles. This simulated behaviour results in a normalised user feedback weighting of:

$$W_{(\text{pre-viewing, clicking result, viewing})}(D_s) = \frac{1.0 + 0.5 + 0.3}{2.5} = 0.68$$

3.4. Simulation-based Evaluation

TABLE 3.1: Weighting of Implicit Features

Event weighting (w_e)	S_1	S_2	S_3	S_4	S_5
$W_{(\text{previewing})}(D_s)$	1.0	–	–	1.0	–
$W_{(\text{clicking result})}(D_s)$	0.5	1.0	1.0	1.0	1.0
$W_{(\text{exploring})}(D_s)$	–	–	0.5	–	1.0
$W_{(\text{viewing})}(D_s)$	(0–1)	(–)	(0–1)	(0–1)	(0–1)

Another example simulation including Sequence S_3 is as follows: We simulate the initial *clicking on result* to start *viewing* for two time cycles. Additionally, we simulate *sliding* to randomly select three shots from the same video. This behaviour will reach a normalised user feedback weighting of:

$$W_{(\text{clicking result, viewing})}(D_s) = \frac{1.0 + 0.2}{2.5} = 0.48$$

and additionally, three random shots from the same video will be taken into account for the next query generation, using the same weighting.

Generate the query for the next iteration

In the third step, we generate the query q_{i+1} for the next iteration. This query is obtained by extracting the x most important terms from the documents involved in interactions in the previous step, by using query expansion techniques. In essence, these are the highest weighted terms of each document. The underlying idea is to simulate users refining their search queries with information from the documents that they interacted with. A new query is formed consisting of the top x weighted terms. Thus,

$$q_{i+1} = \{t_1, \dots, t_x | W_e(t_1) \geq \dots \geq w_e(t_x)\}. \quad (3.1)$$

3.4.2 Parameter Fine Tuning

Before Hypotheses H_1 and H_2 can be studied objectively, the evaluation scheme requires various parameter fine tuning. Parameters are: the initial search query used for retrieval, the number of detected relevant documents x , the percentage of relevant vs. non-relevant results and finally the number of terms y used for query expansion.

Initial Query

The first parameter is the initial query q_0 that is required to start the simulation. An important precondition for the simulation is that the initial set of results provide enough positive results to base the actual query expansion technique on. Aiming to satisfy this condition, the initial search queries were selected manually. The manual queries, which are based on the topic description, consist of one to five terms with a median of 2 and an average of 2.5 terms. The initial retrieval returns an average of 12.9 (median: 9.5) relevant shots out of 100 results over all search tasks.

Number of Documents

The second parameter is the number n of documents that the simulated user is interacting with in each iteration. Figure 3.7 shows the mean average precision of all simulated runs using sequences $S_1 - S_5$ of the Top 5, 10, 15 and 20 relevant results, respectively. Relevance is provided by the given relevance assessment data of each search task.

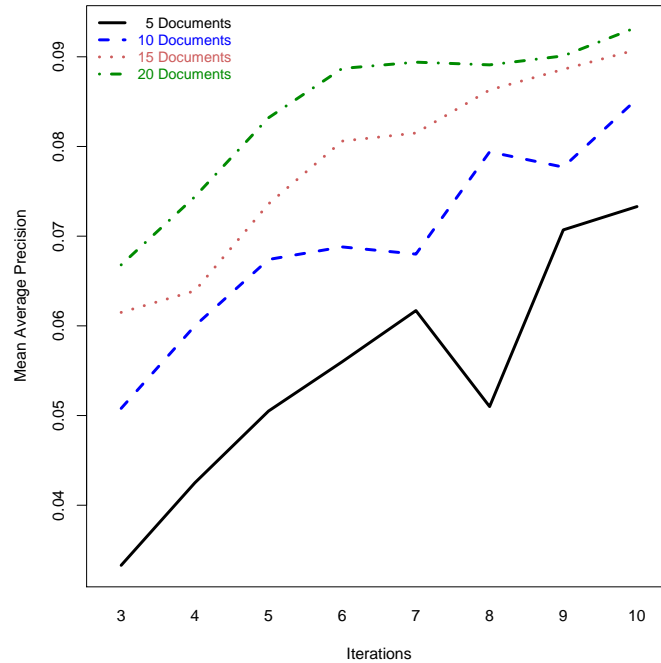


FIGURE 3.7: Mean Average Precision of the number of documents that users interacted with over up to ten iterations

The more relevant documents are used to generate a new search query, the better the mean average precision. However, the more shots are taken into account, the smaller the improvement, compared to runs adding less shots. This derives to the structure of the data set: The shots are associated with only a few keywords (15.89 terms on

average per shot including stop words), hence expanding more results will not result in many new terms. Thus, one can conclude that more relevant shots will return better results. Nevertheless, as the improvement steps get smaller the more results are taken into account, it might be better to perform a query expansion on a smaller set of results, as a user should receive new terms from query expansion earlier rather than later in the interaction process. In our simulation runs, we take the top five results into account. An average of 4.5 relevant shots (median: 5) can be found within the top five search results.

Relevant vs. Non-relevant Results

The third parameter is how many documents n that users interact with should actually be relevant documents.

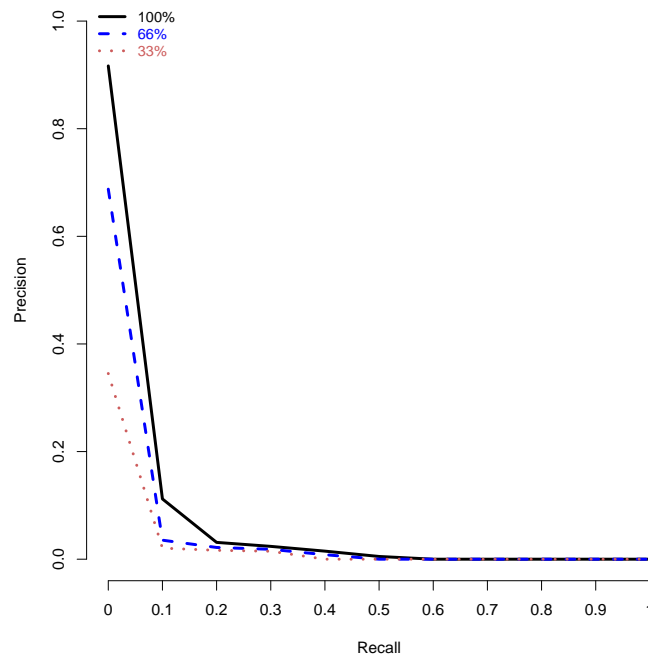


FIGURE 3.8: Precision/Recall of runs with x percent relevant results

Figure 3.8 illustrates the Precision/Recall curve for various percentage of relevant and non-relevant results used for query generation. As expected, the higher the number of non-relevant results taken for query generation, the worse are the retrieval results of subsequent iterations. The reason is obvious: A query generated from terms of non-relevant results will reduce the percentage of relevant terms over each iteration. This phenomena is referred to as query drift, a by-product of automated query expansion [Mitra et al., 1998]. Since the focus of this preliminary analysis is to study the effect of different user interaction scenarios, only relevant results are taken into account. This

allows the best possible results in later iterations. Hence, we simulate a user clicking only on those results which appear to be relevant. This is necessary as otherwise, our system will return too many non-relevant results due to the already weak bounding between key terms and relevance in the TRECVID collection.

Number of Query Terms

The final parameter is the number y of query terms that are used to formulate a new search query. Figure 3.9 shows the mean average precision of retrieval runs using different number of terms y for retrieval.

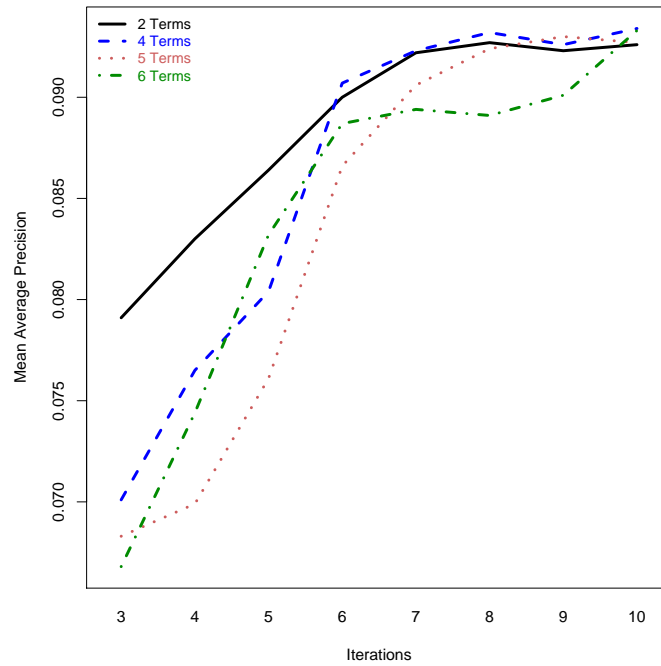


FIGURE 3.9: Number of Terms for Retrieval

The more terms are taken to formulate a new query, the lower is the mean average precision during the first iterations. The reason is that fewer terms are more precise and hence set a stricter focus. Thus, fewer terms will return better retrieval results as they are more focused than more terms. In the simulation, we use a maximum of six terms.

3.4.3 Discussion

After introducing different interaction scenarios in the previous section, we showed in this section how these interaction scenarios can be combined to model simplified user interactions with representative video retrieval interfaces over various iterations I . After

modelling basic user interactions, we analysed the effect of various evaluation parameters. An analysis of these interactions can help to understand the possibility of implicit indicators of relevance. Thus, the main contribution of this section is a simulation-based evaluation scheme that allows evaluating the influence of above introduced implicit indicators of relevance on retrieval performance. The next section discusses the outcome of this simulation.

3.5 Results

The previous section introduced a simplified user model which is employed to evaluate the different user scenarios. In this section, we discuss the results of these simulated runs.

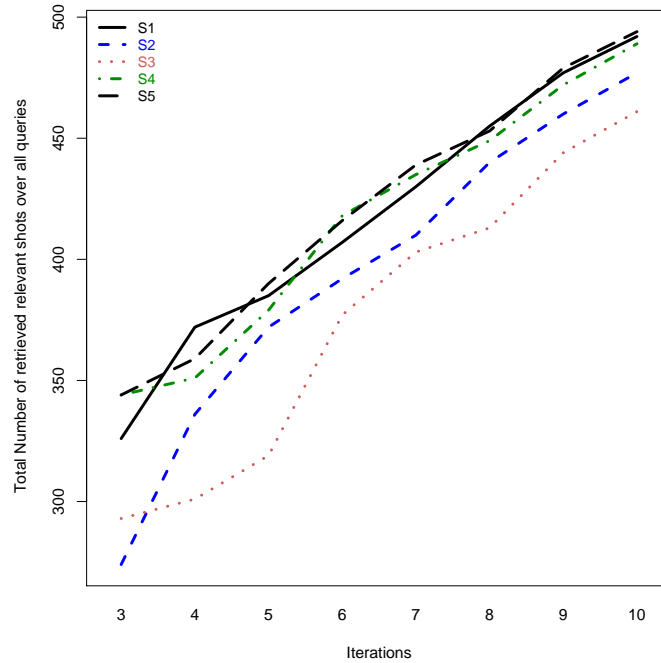


FIGURE 3.10: Total number of retrieved relevant shots over all queries using Sequences S_1 – S_5

Figure 3.10 displays the total number of retrieved relevant shots over all queries over the relevance feedback iterations for the scenarios S_1 – S_5 . As illustrated, the scenarios S_1 , S_4 and S_5 tend to return higher numbers of retrieved relevant shots over all queries than the other two models. Looking at the mean average precision of the test runs (see Figure 3.11), again S_1 , S_4 and S_5 are the most successful models. Comparing both figures, S_3 shows the weakest performance.

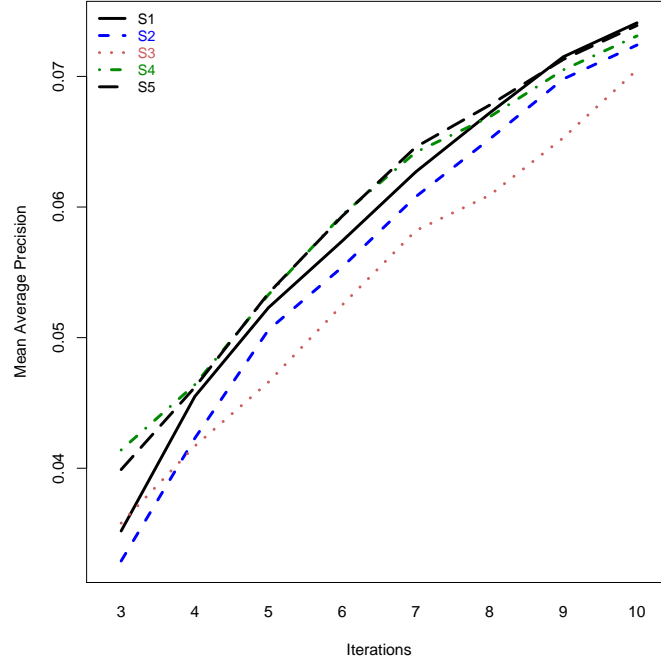


FIGURE 3.11: Mean Average Precision using Sequences S_1 – S_5

Considering the simple setting, a significant analysis of the effect that different low-level event combinations have on the retrieval performance is challenging. Aiming to evaluate both Hypotheses H_1 and H_2 , such detailed analysis is not required though. A preliminary comparison of the overall performance of all scenarios can already shed light on the use of the implicit indicators of relevance.

Scenarios S_2 and S_3 include only few implicit features. As their results return the weakest retrieval results, it suggests that using more implicit indicators can improve retrieval iterations. This would support the first hypothesis (H_1) that implicit relevance feedback can be employed to support interactive video retrieval.

Each scenario returned retrieval results that differ from the other scenarios. As each scenario is the simulation of implicit feedback given by a user, one can conclude that different low-level event combinations directly influenced this difference. This would support the second hypothesis (H_2) that the interpretation and importance of these implicit indicators of relevance depends on the interface context. More precisely, one of the most significant results of the simulation is the similar performance of the systems S_1 and S_4 . S_1 is our basic system while S_4 models the system of [Hopfgartner et al. \[2007\]](#). The only difference between them is that S_4 simulates the browsing through a video. This may indicate that browsing can boost relevant retrieval results. This assumption is supported by the performance of S_5 . It was the most successful model and also includes the simulation of browsing. Thus, S_5 was the only model which in-

cluded the additional simulation of explicit relevance feedback. This correlates with the conclusions taken in the textual domain that the combination of explicit and implicit relevance feedback improves retrieval results. Another interesting result is, that the trend of S_1 over all iterations is very smooth, while the results of the other models are less predictable. S_1 is the only model which does not support browsing or sliding through a video. Thus, the results suggest that using the sliding bar and browsing through a video may be questionable indicators for relevance. News videos are divided into stories. The current story may be of interest for the user. A story can for instance be about a political crisis, while the next story may be about global warming. Taking this into account, shots taken from both sliding and browsing of a video should not be seen as relevant results, as other parts of the news video are less likely to be related to each other. However, the worse results are not a surprise, since all other feedback is based on the ground truth.

Scenario S_1 has the smallest weighting for the playing duration of a video. As it is the best performing model, it casts doubt on the straightforward interpretation of dwell time as an indicator of interest or relevance. This matches the findings of Kelly [2004]. Also, in video retrieval based on key frames, it may obviously take some time (i.e. viewing time) to decide if the segment contains what is required for an answer to the task. The extremely serial nature of videos seems to lead to this conclusion. In fact, the playing duration may be a negative indicator of relevance. The longer a video is played, the more time a user needed to judge the relevance. And the longer needed for that, the less relevant it might be.

3.6 Summary

This chapter has focused on evaluating two hypotheses. The first Hypothesis H_1 states that implicit relevance feedback can be employed to support interactive video retrieval. Therefore, we analysed the influence of implicit features as an indicator for relevance. Based on the interfaces of state-of-the-art adaptive video retrieval systems and the analysis of a small user study, we identified six low-level events which carry the potential to be exploited as implicit indicators of relevance. The second evaluated Hypothesis H_2 was that the interpretation and importance of these implicit indicators depend on the interface context. In order to study both hypotheses, we outlined five different user action sequences $S_1 - S_5$ which include different combinations of these events. Based on these scenarios, we ran a simulation-based user study to see, if the different combinations of features can have an influence on retrieval results. In this evaluation methodology, we assume users interacting with video retrieval systems following the introduced in-

3.6. Summary

teraction patterns. The simulation outcome illustrates different performances for each user interface scenario. As the various user interaction scenario simulations perform differently, we conclude that implicit features *do* have an influence on interactive video retrieval results.

In summary, it can be concluded that both Hypotheses H_1 and H_2 have been supported by the results of this simulation-based user study. This simulated methodology is a pre-implementation method though. Given the numerous combinations of features and interface scenarios, we select an appropriate number of them. This will give a further opportunity to develop appropriate systems and subsequent user-centred evaluation. The real effect of a video retrieval system only can be measured by user experiments. The presented approach, however, provides a mechanism to benchmark a number of possible models before it reaches implementation. In the next chapter, we introduce a user-centred evaluation of a recommendation approach that exploits implicit relevance feedback. Its result should gain a deeper insight into the use of implicit indicators of relevance.

– *One cannot not communi-
cate.*

Paul Watzlawick, 1967

4

Exploiting Community-Based Relevance Feedback

Having established a simulation-based evaluation scheme to evaluate the role of implicit relevance feedback in the video retrieval domain, this chapter introduces an evaluative user study aiming to further examine this role. We evaluate whether implicit relevance feedback provided by a community of users can effectively be exploited to recommend relevant videos. Section 4.1 introduces the topic. In Section 4.2, we introduce our approach of employing the implicit indicators of relevance. Section 4.3 introduces a video retrieval system which is used in the evaluation. The experiment is outlined in Section 4.4, results are presented in Section 4.5. Section 4.6 summarises the chapter.

4.1 Introduction

In the previous chapter, we studied whether implicit relevance feedback can be employed for short term user profiling in the video retrieval domain. Therefore, we first identified the most common implicit indicators of relevance by analysing representative video retrieval interfaces. We have further shown that design of these interfaces directly influences users' interaction patterns. We hypothesised that the interpretation and importance of the implicit indicators depends on the interface context. We verified this hypothesis by simulating users interacting with these interfaces over various iterations. The outcome of this simulation-based evaluation tentatively support this hypothesis.

As argued, however, the simulation of user interactions should only be seen as a preliminary evaluation method. In this chapter, we therefore evaluate the application of implicit relevance feedback in the video retrieval domain by performing a user-centred evaluation.

Developing the assertions of the previous chapter further, we hypothesise in this chapter that not only can implicit relevance feedback be used to improve retrieval performances by adapting a user's search query, but also to generate appropriate recommendations. Recommendations allow users to discover further documents that they have not seen before and can thus help bridging the Semantic Gap that makes video retrieval so challenging. Recommendation techniques based on implicit low-level events do not require users to alter their normal behaviour, while all of the actions that users carry out can be used to improve their retrieval results.

Considering that systems such as YouTube or Dailymotion enable many users to search for similar search topics, we further argue that collaborative or community-based relevance feedback can be employed to provide corresponding recommendations. Many of the earliest collaborative techniques emerged online in the 1990s [Goldberg et al., 1992; Resnick et al., 1994; Shardanand and Maes, 1995] and focused on the notion of collaborative filtering. Collaborative filtering was first developed in the Tapestry system to recommend e-mails to users of online newsgroups [Goldberg et al., 1992]. Collaborative filtering aims to group users with similar interests, with a view to treating them similarly in the future. So, if two users have consistently liked or disliked the same resources, then chances are that they will like or dislike future resources of that type. Since those early days collaborative or community-based methods have evolved and been used to aid browsing [Wexelblat and Maes, 1999], e-learning [Freyne et al., 2007] and in collaborative search engines [Smyth et al., 2004]. More recently there has also been some recent initial research into carrying out collaborative video search [Adcock et al., 2007]. This work, however, concentrated on two users carrying out a search simultaneously rather than using the implicit interactions from previous searches to improve future searches. We hypothesise that there are a number of potential benefits of exploiting implicit relevance feedback from multiple users that have been searching for similar topics at different times. We test this assertion by introducing a graph-based model that utilises the implicit actions of previous user searches. First of all, we argue that recommendations that have been determined by implicit means will improve retrieval performance. Therefore, we hypothesise that the performance of the users of the system, in terms of precision of retrieved videos, will improve with the use of recommendations based on feedback. Moreover, we assert that the users will be able to explore a collection to a greater extent, and also discover aspects of a topic that

4.2. A Graph-Based Approach for Capturing Implicit Relevance Feedback

they may not have considered since they will be presented with new documents they have not seen before. Consequently, we assume that users will be more satisfied with the system that provides feedback, and also be more satisfied with the results of their search. Summarising, this chapter aims to study the following hypotheses:

H₁: Implicit relevance feedback can be employed to support interactive video retrieval.

H₂: The interpretation and importance of the implicit indicators of relevance depend on the interface context.

H₃: Implicit relevance feedback can be employed to recommend relevant video documents.

H₄: Users will be able to explore a collection to a greater extent, and also discover aspects of the topic that they may not have considered, when implicit relevance feedback is used to recommend related documents.

H₅: Users will be satisfied with using a system that provides relevant recommendations by exploiting implicit relevance feedback.

In Section 4.2, we therefore introduce a graph-based model that utilises implicit actions involved in previous user searches. The model can provide recommendations to support users in completing their search tasks. Two systems, introduced in Section 4.3, are compared. The first system is a baseline system that provides no recommendations. The second system is a system that provides recommendations based on our model of implicit user actions. In Section 4.4, both systems and their respective performances are evaluated both qualitatively and quantitatively. Results are shown in Section 4.5 and discussed in Section 4.6.

The research which is presented in this chapter has been published in [Vallet et al., 2008a; Urban et al., 2006a; Hopfgartner et al., 2008d,c].

4.2 A Graph-Based Approach for Capturing Implicit Relevance Feedback

As highlighted in the previous chapter, various implicit indicators can be used to infer relevance of a document. For our recommendation model based on user actions, there are two main desired properties of the model for action information storage. The first property is the representation of all of the user interactions with the system, including

4.2. A Graph-Based Approach for Capturing Implicit Relevance Feedback

the search trails for each interaction. This allows us to fully exploit all of the interactions which have been introduced in Section 3.2 to provide richer recommendations. The second property is the aggregation of implicit information from multiple sessions and users into a single representation, thus facilitating the analysis and exploitation of past implicit information. To achieve these properties we opt for a graph-based representation of the users' implicit information. We adopt the concept of trails from White et al. [2007], except we do not limit the possible recommended documents to those documents that are at the end of the search trail. We believe that during an interactive search the documents that most of the users with similar interaction sequences interacted with, are the documents that could be most relevant for recommendation, not just the final document in the search trail. Similar to Craswell and Szummer [2007], our approach represents queries and documents in the same graph, however we represent the whole interaction sequence, unlike their approach where the clicked documents are linked directly to the query node. Once again we want to recommend potentially important documents that are part of the interaction sequence and not just the final document of this interaction. Considering the specific low-level events which are unique to multimedia search engines (introduced in Section 3.2), our representation exploits a greater range of user interactions in comparison with other approaches [Craswell and Szummer, 2007; White et al., 2007]. This produces a more complete representation of a wide range of user actions that may facilitate better recommendations. These properties and this approach result in two graph-based representations of user actions. The first representation utilises a Labelled Directed Multigraph (LDM) for the detailed and full representation of implicit information. The second graph is a Weighted Directed Graph (WDG), which interprets the information in the LDM and represents it in such a way that is exploitable for a recommendation algorithm. The recommendations that are provided are based on three different techniques based on the WDG. The two graph representation techniques and the recommendation techniques are described in detail in the following sections.

4.2.1 Labelled Directed Multigraph

A user session s can be represented as a set of queries Q_s , which were input by the user u , and a set of multimedia documents D_s the users interacted with during the search session. Queries and documents are represented as nodes $N_s = \{Q_s \cup D_s\}$ of our graph representation, $G_s = (N_s, A_s)$. The interactions of the user during the search session are represented as a set of action arcs $A_s(G) = \{n_i, n_j, a, u, t\}$. Each action arc indicates that, at a time t , the user u performed an action of type a that leads the user from the query

4.2. A Graph-Based Approach for Capturing Implicit Relevance Feedback

or document node n_i to node $n_j, n_i, n_j \in N_s$. Note that n_j is the object of the action and that actions can be reflexive. For instance, when a user clicked to view a video and then navigate through it. Action types depend on the kind of actions recorded by the implicit feedback system. In our system we consider all low-level feedback events which have been introduced in Section 3.2, namely start playing a video (viewing), navigating through a video (sliding), highlighting a video to get additional metadata (exploring) and selecting a video (clicking result). Links can contain extra associated metadata, as type specific attributes, e.g. length of play in a play type action. The graph is multi-linked, as different actions can have the same source and destination nodes. The session graph $G_s = (N_s, A_s)$ will then be constructed by all the accessed nodes and linking actions, and will represent the whole interaction process for the user's session s . Finally, all session-based graphs can be aggregated into a single graph $G = G(N, A)$, $N = \cup_s N_s$, $A = \cup_s A_s$ which represents the overall pool of implicit information. Subsequently, all of the nodes from the individual graphs are mapped to one large graph, and then all of the action edges are mapped onto the same graph. This graph may not be fully connected, as it is possible, for instance, that users selected different paths through the data or entered a query and took no further actions, etc. While the LDM gives a detailed representation of user interaction with the collection, it is extremely difficult to provide recommendations. The multiple links make the graph extremely complex. In addition, all of the actions are weighted equally. This is not always a true representation; some actions may be more important than others and should be weighted differently.

4.2.2 Weighted Directed Graph

In order to allow the recommendation algorithm to exploit the LDM representation of user actions, we convert the LDM to a WDG by collapsing all links interconnecting two nodes into one single weighted edge. This process is carried out as follows. Given the detailed LDM graph of a session s , $G_s = (N_s, A_s)$, we compute its correspondent weighted graph $G_s = (N_s, W_s)$. Links $W_s = \{n_i, n_j, w_s\}$ indicate that at least one action lead the user from the query or document node n_i to n_j . The weight value w_s represents the probability that node n_j , was relevant to the user for the given session, this value is either given explicitly by the user, or calculated by means of the implicit evidence obtained from the interactions of the user with that node:

$$W_s(n_i, n_j) = \begin{cases} 1, & \text{iff relevance feedback type for } n_j = \text{explicit} \\ -1, & \text{iff non-relevance feedback type for } n_j = \text{explicit} \\ lr(n_j) \in [0, 1], & \text{otherwise (i.e. implicit relevance)} \end{cases}$$

In the case that there is only implicit evidence for a node n , the probability value is given by the local relevance $lr(n)$. $lr(n)$ returns a value between 0 and 1 that approximates a probability that node n was relevant to the user given the different interactions that the user had with the node. For instance if the user opened a video and played it for the whole of its duration, this can be enough evidence that the video has a high chance of being relevant to the user. Following this idea, the local relevance function is defined as $lr(n) = 1 - \frac{1}{x(n)}$, where $x(n)$ is the total of added weights associated to each type of action in which node n is an object of. This subset of actions is defined as $A_s(G_s, n) = \{n_i, n_j a, u, t \mid n_j = n\}, n \in N_s$. These weights are natural positive values returned by a function $f(a) : A \rightarrow N$, which maps each type of action to a number. These weights are higher for an action that is understood to give more evidence of relevance to the user. In this way, $lr(n)$ is closer to 1 as more actions are observed that involve n and the higher the associated weight given to each action type. In our weighting model some of the implicit actions are weighted nearly as highly as explicit feedback. The accumulation of implicit relevance weights can thus be calculated as $x(n) = \sum_{a \in A_s(G_s, n)} f(a)$. Table 4.1 shows an example of function f , used during our evaluation process: all of these low-level events are part of the system described in Section 4.3. This system considers the following actions which have been determined in Section 3.2: (1) playing a video during a given interval of time (Viewing); (2) clicking a search result in order to view its contents (Clicking result); (3) navigating through the contents of a video (Sliding); (4) browsing to the next or previous video key frame (Browsing R/L) and (5) tooltipping a search result by leaving the mouse pointer over the search result (Pre-viewing). Table 4.1 shows the weights that are assigned to these low-level events for our study. The weights are based on previous work by Hopfgartner et al. [2007].

Figure 4.1 shows an example of LDM and its correspondent WDG for a given session. Similarly to the detailed LDM graph, the session-based WDGs can be aggregated into a single overall graph $G = (N, W)$, which will be called the implicit relevance pool, as it collects all the implicit relevance evidence of all users across all sessions. The nodes of the implicit pool are all the nodes involved in any past interaction $N = \cup_s N_s$, whereas the weighted links combine all of the session-based values. In our approach we opted for a simple aggregation of these probabilities, $W = \{n_i, n_j, w\}, w = \sum_s w_s$ since

4.2. A Graph-Based Approach for Capturing Implicit Relevance Feedback

TABLE 4.1: Values for function $f(a)$ used during the experiment

Action a	$f(a)$
Viewing	3
Clicking result/Playing	10
Sliding/Navigating	2
Browsing R/L	2
Previewing/ToolTip	1

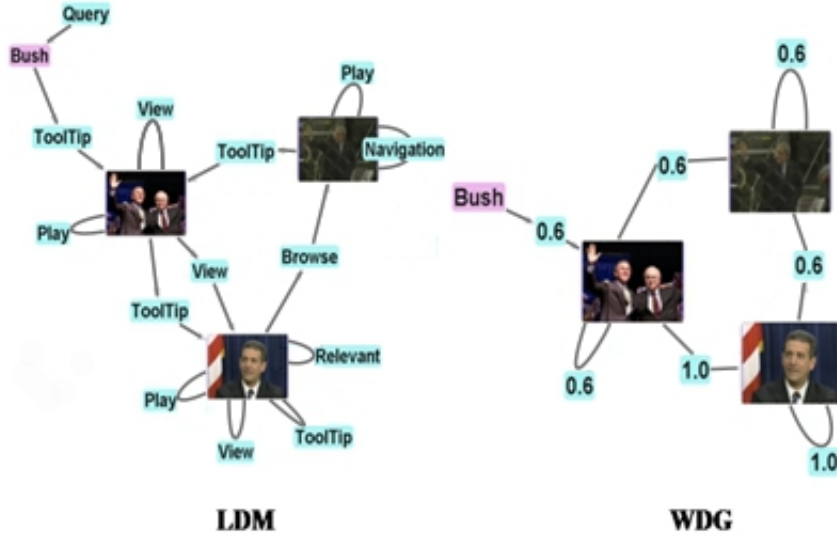


FIGURE 4.1: Correspondence between the LDM (*left*) and WDG (*right*) models

this treats all probabilities equally. Each link represents the overall implicit (or explicit, if available) relevance that all users, whose actions lead from node n_i to n_j , gave to node n_j . Figure 4.2 shows an example of the implicit relevance pool.

4.2.3 Implicit Relevance Pool Recommendation Techniques

In our system we recommend both queries and documents to the users. These recommendations are based on the status of the current user session. As the user interacts with the system, a session-based WDG is constructed. This graph is the basis of the recommendation algorithm which has three components; each component uses the implicit relevance pool in order to retrieve similar nodes that were somehow relevant to other users. The first two components are neighbourhood based. A neighbourhood approach is a way of obtaining related nodes; we define the node neighbourhood of a given node n , as the nodes that are within a distance D_{MAX} of n , without taking the link direction-

4.2. A Graph-Based Approach for Capturing Implicit Relevance Feedback

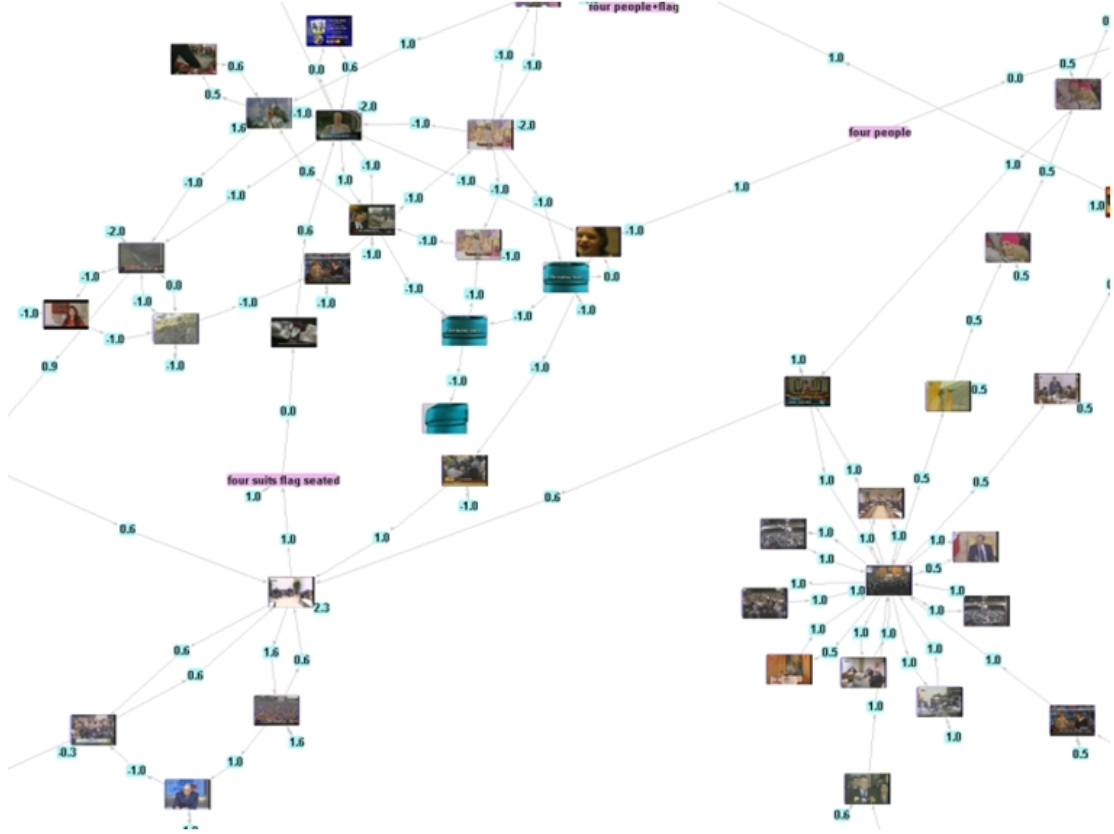


FIGURE 4.2: Graph illustrating implicit relevance pool

ality into consideration. These nodes are somehow related to n by the actions of the users, either because the users interacted with n after interacting with the neighbour nodes, or because they are the nodes the user interacted with after interacting with n . More formally as a way of obtaining related nodes, we define the node neighbourhood of a given node n as:

$$NH(n) = \{m \in N \mid \delta(n, m) \leq D_{MAX}\}$$

where $\delta(n, m)$ is the shortest path distance between nodes n and m , and D_{MAX} is the maximum distance in order to take into consideration a node as a neighbour. The best performing setting for this value, in our experiments, was $D_{MAX} = 3$.

Using the properties derived from the implicit relevance pool, we can calculate the overall relevance value for a given node. This value indicates the aggregation of implicit relevance that users gave historically to n , when n was involved with the users' interactions. Given all the incident weighted links of n , defined by the subset $W_s(G_s, n) = \{n_i, n_j, w \mid n_j = n\}, n \in N_s$ the overall relevance value for n is calculated as follows:

$$or(n) = \sum_{w \in w_s(G_s, n)} w$$

Given the ongoing user session s , and the implicit relevance pool we can then define the node recommendation value as:

$$nh(n, N_s) = \sum_{n_i \in N_s, n \in NH(n_i)} lr(n_i) \cdot or(n)$$

where $lr(n_i)$ is the local relevance computed for the current session of the user, so that the relevance of the node to the current session is taken into consideration.

Exploiting this node recommendation value, we can then determine the recommendations for all *queries* and the recommendations for all *documents* within the user's current search session, i.e. the highest weighted neighboured nodes. The last recommendation component is based on the user's interaction sequence. The interaction sequence recommendation approach tries to take into consideration the interaction process of the user, with the scope of recommending those nodes that are following this sequence of interactions. For instance, if a user has opened a video of news highlights, the recommendation could contain the more in-depth stories that previous users found interesting to view next. Therefore, we identify all nodes in the pool that can be reached from the nodes of the user's current session by following any path within the pool.

In a final step, we obtain the three recommendation lists from each recommendation component and merge them into a single final recommendation lists. Assuming that recommendations can be ranked based on their relevance to the user's current interest, we use a rank-based aggregation approach where the scores of the final recommendations are the sum of the rank-based normalised scores of each of the recommendation list, i.e. using a score $\frac{1}{r(n)}$ where $r(n)$ is the position of n in the recommended list. This way, we can guarantee that no recommendation source is favoured, since the recommendations are sorted based on their ranking only. The final list is then split into recommended queries and recommended documents; these are then presented to the user.

4.3 System Description

In order to evaluate our hypotheses, our implicit feedback approach has been implemented in an interactive video retrieval system. This allows us to have actual end users test our system and approach. The shots in our collection were indexed using Terrier based on ASR transcript and machine translation output. The Okapi BM25 retrieval model was used to rank retrieval results. In addition to the ranked list of search results,

4.3. System Description

the system provides users with additional recommendations of video shots that might match their search criteria based on our recommendation graph (see Section 4.2 for details on the recommendation graph).

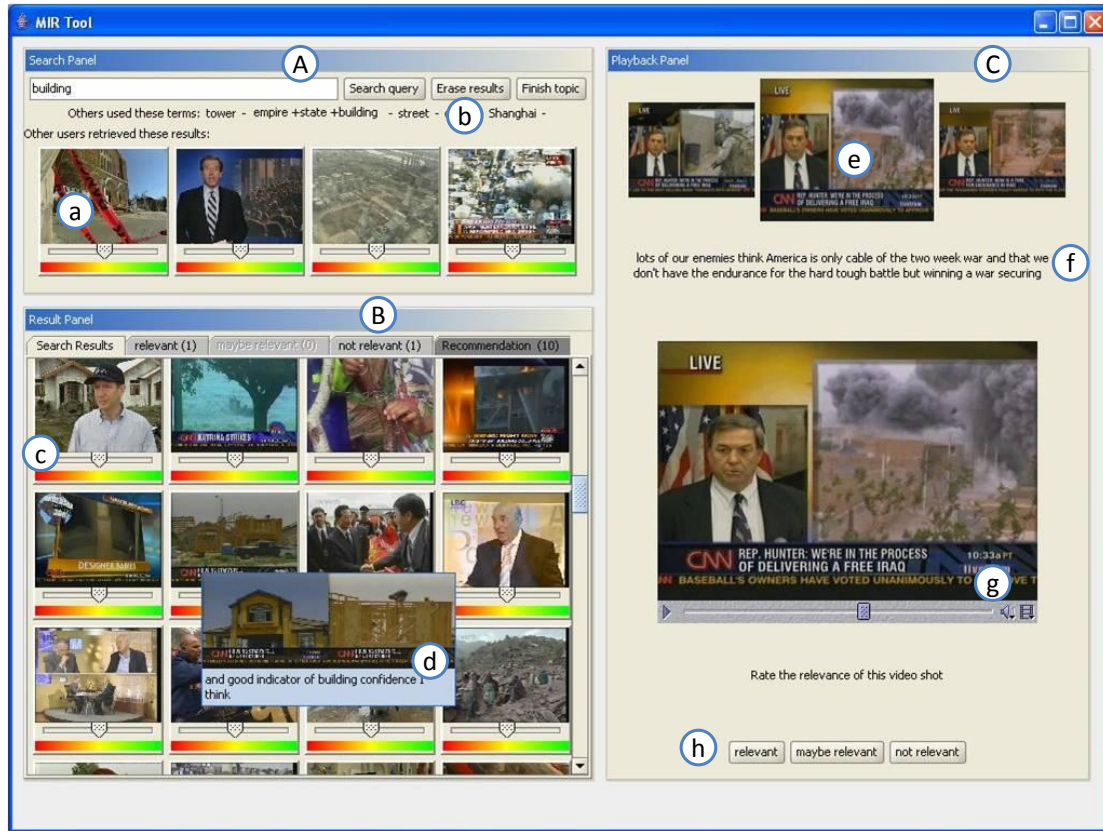


FIGURE 4.3: Interface of the video retrieval system

Figure 4.3 shows a screen shot of the recommendation system. The design of the interface is based on the interfaces that have been introduced in Section 2.1.5. It can be divided into three main panels: the search panel (A), the result panel (B) and the playback panel (C). The most common feature of all interfaces is the search panel (A), where users formulate and carry out their searches. Users can enter a text-based query in the search panel (A) to begin their search. The users are presented with text-based recommendations for search queries that they can use to enhance their search (b). The users are also presented with recommendations of video shots that might match their search criteria (a). Each recommendation is only presented once, but may be retrieved by the user at a later stage if they wish to do so. The result panel is where users can view the search results (B). This panel is divided into five tabs, the results for the current search, a list of results that the user has marked as relevant, a list of results that the user has marked as maybe being relevant, a list of results that the user has marked

as irrelevant and a list of recommendations that the user has been presented with previously. Users can mark results in these tabs as being relevant by using a sliding bar (c). The advantage of this technique is that users can bookmark video shots as “maybe relevant”, even though they are not sure yet whether the shot really shows what they are looking for. Additional information about each video shot can be retrieved by hovering the mouse cursor over a video key frame, that key frame will be highlighted, along with neighbouring key frames and any text associated with the highlighted key frame (d). Similar to those interfaces that have been introduced in Section 2.1.5, retrieval results are displayed in ranked order. The playback panel (C) is for viewing video shots (g). As a video is playing it is possible to view the current key frame for that shot (e), any text associated with that key frame (f) and the neighbouring key frames. Users can play, pause, stop and can navigate through the video as they can on a normal media player, and also make relevance judgements about the key frame (h). Some of these tools in the interface allow users of the system to provide explicit and implicit feedback, which is then used to provide recommendations to future users. Explicit feedback is given by users by marking video shots as being either relevant or irrelevant (c, h). Implicit feedback is given by users playing a video (g), highlighting a video key frame (d), navigating through video key frames (e) and selecting a video key frame (e). Both search panel (A) and result panel (B) are visible at all times, allowing the user to formulate new search queries and/or preview other retrieval results while playing a video shot.

In order to provide a comparison to our recommendation system, we also implemented a baseline system that provides no recommendations to users. The baseline system has previously been used for the interactive search task track at TRECVID 2006 [Urban et al., 2006a], the performance of this system was average when compared with other systems at TRECVID that year. A tooltip feature which shows neighbouring key frames and the transcript of a shot was added to this system to improve its performance. Overall the only difference between the baseline and the recommendation system is the provision of key frame recommendations (a).

4.4 Experimental Methodology

4.4.1 Collection and Tasks

In order to determine the effects of implicit feedback, users were required to carry out a number of video search tasks based on the TRECVID 2006 collection and tasks. For our evaluation we focus on the interactive search tasks that involve the use of low-level content-based search techniques and feedback from users of the video search system.

For the interactive search task users are given a specific query and a maximum of 15 min to find shots relevant to that query. Voorhees [2005] argues that at least 24 different tasks are required to gather statistical significant results from such user experiments. Therefore, datasets such as TRECVID consist of at least 24 different search tasks. However, the goals of this evaluation were not the same as within TRECVID, which aims for comparing the effectiveness of different systems. In order to study system specific research questions with reasonable cost and effort, a well established approach (e.g. [Halvey et al., 2009a; Villa et al., 2008b]) is to limit the number of tasks that the users carry out. For this evaluation we follow this limitation by considering a subset of the TRECVID tasks for evaluation. We opt for four tasks for which the average precision in the 2006 TRECVID workshop was the worst. In essence these are the most *difficult* tasks. The four tasks were chosen as in general these are tasks for which there are less relevant documents. Indeed the mean average precision (MAP) values show that it is extremely difficult to find these documents. We argue that any improvement which may be gained on these difficult tasks with few documents will be reflected on less difficult tasks with larger numbers of relevant documents. The same cannot be said about the gains made for easier tasks being borne out in more difficult tasks. Moreover, due to the difficult nature of these topics, different users had to use a different search query, which ensures that users do not just follow other users' search trails (this is shown in subsequent sections). As can be seen below there were very few relevant shots in the collection for these tasks, 98 shots out of 79,848 shots for one of the tasks. In addition to this, not all of the relevant shots have text associated with them. As the most popular form of search is search by textual query [Christel, 2007a], finding these shots becomes even more difficult. The four tasks that were used for this evaluation were:

1. Find shots with a view of one or more tall buildings (more than 4 stories) and the top story visible (142 relevant shots, 53 with associated text)
2. Find shots with one or more soldiers, police, or guards escorting a prisoner (204 relevant shots, 106 with associated text)
3. Find shots of a group including at least four people dressed in suits, seated, and with at least one flag (446 relevant shots, 287 with associated text)
4. Find shots of a greeting by at least one kiss on the cheek (98 relevant shots, 74 with associated text).

The users were given the topic and a maximum of 15 min to find shots relevant to the topic. The users could only carry out text-based queries, as this is the normal method

of search in most online and desktop video retrieval systems and also the most popular search method at TRECVID [Christel, 2007a]. The shots that were marked as relevant were then compared with the ground truth in the TRECVID collection.

4.4.2 Experimental Design

For our evaluation we adopted a 2-searcher-by-2-topic Latin Square design. Each participant carried out two tasks using the baseline system, and two tasks using the recommendation system. The order of system usage was varied as was the order of the tasks; this was to avoid any order effect associated with the tasks or with the systems. To determine the effect of adding more implicit actions to the implicit pool, participants in the experiment were placed in groups of four. For each group, the recommendation system used the implicit feedback from all of the previous users. At the beginning of the evaluation there was no pool of implicit actions, therefore the first group of four users received no recommendations; their interactions formed the training set for the initial evaluations. Using this experimental model we can evaluate the effect of the implicit feedback within a group of participants, and also the effect of additional implicit feedback across the entire group of participants. In addition to this, the ground truth provided in the TRECVID 2006 collection allowed us to carry out analyses that we may not have been able to do with other collections that do not have corresponding ground truth data. Each participant was given 5 min training on each system and carried out a training task with each system. These training tasks were the tasks for which participants had performed the best at TRECVID 2006. For each participant their interaction with the system was logged, the videos they marked as relevant were stored and they also filled out a number of questionnaires at different stages of the experiment. The documents and questionnaires used for this experiment can be found in Appendix A.

4.4.3 Participants

24 participants took part in our evaluation and interacted with our two systems. The participants were mostly postgraduate students and research assistants. The participants consisted of 18 males and 6 females with an average age of 25.2 years (median: 24.5) and an advanced proficiency with English. Students were paid a sum of £10 for their participation in the experiment. Prior to the experiment the participants were asked to fill out a questionnaire so that we could ascertain their proficiency with and experience of dealing with multimedia. We also asked participants about their knowledge of news stories, as the video collection which the participants would be dealing with consists of mainly news videos. It transpired that the participants follow news stories/events once

or twice a week and also watch news stories online. The majority of participants deal with multimedia regularly (once or twice a day) and are quite familiar with creating multimedia data (images, videos). The participants also had a great deal of experience of searching for various types of multimedia. These activities were mainly carried out online, with Flickr, Google or YouTube being cited as the most commonly used online services. The most common search strategy that users mentioned was searching for data by using initial keywords and then adapting the query terms to narrow down the search results based on the initial results received. Using the recommendations provided by some of these services was also mentioned by a number of users. Although the participants often searched for multimedia data, they stated that they rarely use multimedia management tools to organise their personal multimedia collection. The most common practise among the participants is creating directories and files on their own personal computer. Categorising videos and images according to the content and time when this data was produced, is the most popular method of managing media. However, when asked how a system could support their own search strategy, many participants mentioned that it would be helpful to sort or retrieve multimedia based on their semantic content. The following section outlines the results of our evaluation.

4.5 Results

4.5.1 Task Performance

As we were using the TREC Vid collection and tasks, we were able to calculate standard evaluation values for all of the tasks. Figure 4.4 shows the P@N for the baseline and recommendation systems for varying values of N.

Figure 4.5 shows the MAP for baseline and recommendation systems for different groups of users. Each group of four users also had additional feedback from previous participants, which the previous group of four users did not have. It shows the effectiveness of the recommendation technique when compared with the introduced baseline.

Figure 4.6 shows the average time in seconds that it takes a user to find the first relevant shot for both the baseline and the recommender systems. The results of this figure indicate that the system that uses recommendations outperforms the baseline system in terms of precision. It can be seen quite clearly from Figure 4.4 that the shots returned by the recommendation system have a much higher precision over the first 5–30 shots than the baseline system. We verified that the difference between the two P@N values for values of N between 5 and 100 was statistically significant using a pair wise t -test ($p = 0.0214$, $t = 3.3045$). Over the next 100–2000 shots the difference is negligible.

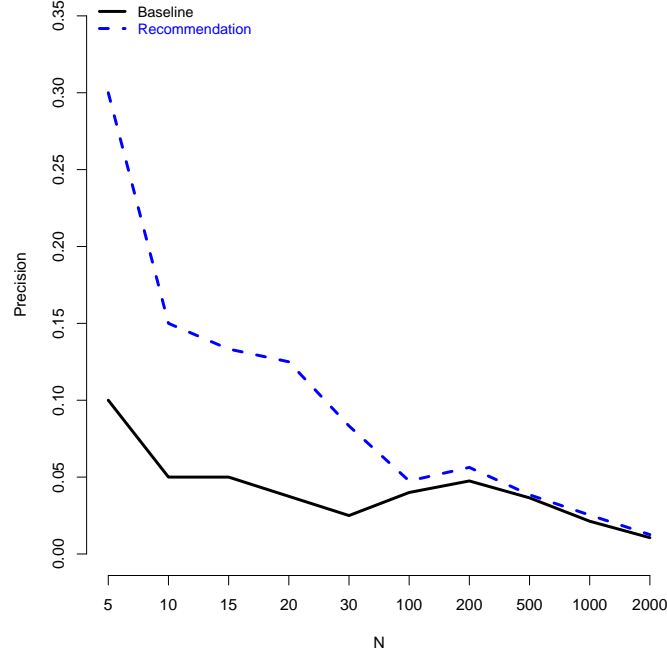


FIGURE 4.4: P@N for the baseline and recommendation systems for varying values of N

However, it is unlikely that a user would view that number of shots; given that in total our 24 participants viewed 3034 shots (see Table 4.2), in the entire trial, 24 hours of video viewing. This demonstrates that the use of the implicit feedback can improve the retrieval results of the system, hence supporting both Hypotheses H_1 and H_2 that “implicit relevance feedback can be employed to support interactive video retrieval” and that “the interpretation and importance of the implicit indicators of relevance depend on the interface context”.

Figure 4.5 shows that the MAP values of the shots the participants selected using the recommendation system are higher than the MAP values of the shots that the participants selected using the baseline system. We verified that the difference between the two sets of results was statistically significant using a pair wise t -test ($p = 0.0028$, $t = 6.5623$). The general trend is that the MAP values of the shots found using the recommendation system is increasing with the amount of training data that is used to propagate the graph-based model. There is a slight dip in one group; however, this may be due to the small sample groups that we are using.

However, these findings are not quite borne out by the recall values for the tasks. In general the recall is low for all of the systems for all of the tasks at TREC Vid 2006; the main focus is on the precision values. While recall is an important aspect we argue that it is more important that the users found accurate results and that they perceived that they had explored the collection, as they had found a heterogeneous set of results. While

4.5. Results

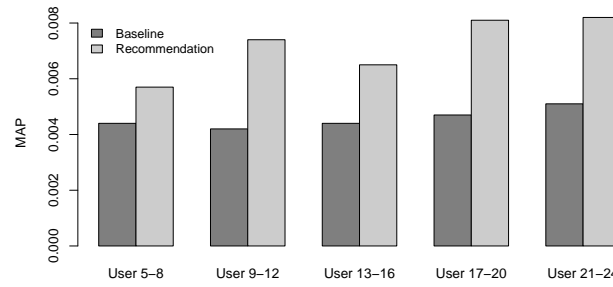


FIGURE 4.5: Mean average precision (MAP) for baseline and recommendation systems for different groups of user

the results in Figures 4.4 and 4.5 show that the users are presented with more accurate results and find more accurate results, this is not telling the full story. In a number of scenarios users will just want to find just one result to satisfy their information need. Figure 4.6 shows that for three of the four tasks the users using the recommendation system find their first relevant result more quickly than the users using the baseline system. The one task for which the baseline system outperforms the recommendation system is due to the actions of two users who did not use the recommendations. We do not know why these two users did not use the recommendations, as they did utilise the recommendations for the other task which they carried out using the recommendation system. A closer examination of the users who did use the recommendations found that three users found relevant shots in less than 1 min, none of the users using the baseline system managed to find relevant shots in less than a minute. Overall the difference in values is not statistically significant.

The results presented so far have shown that users do achieve more accurate results using the system that provides recommendations. We measured P@N and MAP values; it has been shown that the recommendation system outperforms the baseline system, and that this difference is statistically significant. It can be seen that overall the system that is providing recommendations is returning more accurate results to the user. As a result of this, the users are interacting with more relevant videos and find more accurate results. In addition to this, users are finding relevant videos more quickly using the recommendation system (see Figure 4.6). This demonstrates the validity of Hypothesis H_4 , since the performance of the users of the system, in terms of precision of retrieved videos, has improved with the use of recommendations based on implicit feedback. In the following subsection we will discuss user exploration of the collection in more detail.

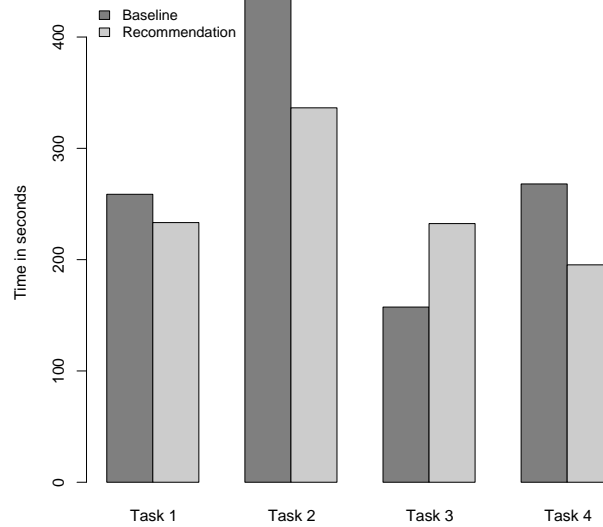


FIGURE 4.6: Average time in seconds to find first relevant shot for baseline and recommendation systems

4.5.2 User Exploration

User Interactions

As was outlined in the system description (Section 4.3) there are a number of ways that the participants could interact with the system. Once a participant enters an initial query in our system, all of the available tools may be used to browse or search the video collections. We begin our investigation of user exploration by briefly analysing these interactions.

Table 4.2 outlines how many times each low-level event available was used across the entire experimental group. During the experiments, the participants entered 1083 queries; many of these queries were unique. This indicates that the participants took a number of different approaches to the tasks, indicating that their actions were not determined by carrying out the same tasks. The figures in Table 4.2 also show that participants play shots quite often. However, if a video shot is selected then it plays automatically in our system. This makes it more difficult to determine whether participants are playing the videos for additional information or if the system is doing so automatically. To compensate for this we only count a play action if a video plays for more than 3 seconds. Another feature that was widely used in our system was the tooltip feature. The tooltip highlighting functionality allowed the users to view neighbouring

4.5. Results

TABLE 4.2: Event type and the number of occurrences during the experiment

Event Type	Occurrences
Query	1083
Mark relevant	1343
Mark maybe relevant	176
Mark not relevant	922
View	3034
Play (for more than 3 s)	7598
Browse key frames	814
Navigate within a video	3794
Tooltip	4795
Total actions	23559

key frames and associated text when moving the mouse over one key frame. This meant that the participants could get context and a feel for the shot without actually having to play that shot.

Analysis of Interaction Graph

In order to gain further insight into the users' interactions a number of different aspects of the interaction graph were analysed. In particular we were interested in investigating changes in the graph structure as additional users used the system. These aspects include the number of nodes, the number of unique queries and the number of links that were present in the graph.

TABLE 4.3: Number of graph elements in graph after each group of four users

Users	# nodes (%)	# queries (%)	# edges (%)	Total graph elements (%)
1–4	1001 (28.31)	115 (18.51)	2505 (23.09)	3621 (24.13)
1–8	1752 (49.56)	258 (41.54)	4645 (42.81)	6655 (44.35)
1–12	2488 (70.38)	388 (62.48)	7013 (64.63)	9989 (66.57)
1–16	3009 (85.12)	452 (72.79)	8463 (78)	11924 (79.46)
1–20	3313 (93.72)	550 (88.57)	9868 (90.95)	13731 (91.5)
1–24	3535 (100)	621 (100)	10850 (100)	15006 (100)

Table 4.3 shows the results of this analysis. It can be seen in Table 4.3 that the number of new interactions increases as the number of participants also increases. The majority of nodes in the graph are video shots (apart from query nodes), as the number of participants increases so does the number of unique shots that have been viewed. On further investigation of the graph and logs it was found that, overall, 49% of documents

selected by users 1–12 were selected at least by one user in 13–24. Users 1–12 clicked 1050 unique documents, whereas users 13–24 clicked 596 unique documents. Also, users 1–12 produced 1737 clicks, whereas users 13–24 produced 1024. This can be interpreted as users 13–24 were satisfied more quickly than users 1–12. It was also found that the number of unique queries also increases with the additional users. These results give an indication that later participants are not just using the recommendations to mark relevant videos, but also interacting with further new and unique shots.

Top Retrieved Videos

Aiming to analyse the reliability of the implicit relevance feedback, we were interested to see whether the nodes in the graph, i.e. the nodes that users provided implicit relevance feedback on, are really relevant. Exploiting the ground truth data, we therefore computed the probability of being relevant of all shots with the same edge weight e , i.e.

$$P(R|s) = \frac{\# \text{ of relevant stories with edge weight } e}{\# \text{ of relevant stories}}. \quad (4.1)$$

Figure 4.7 plots the probability against this relevance value that is assigned to that document from our graph representation.

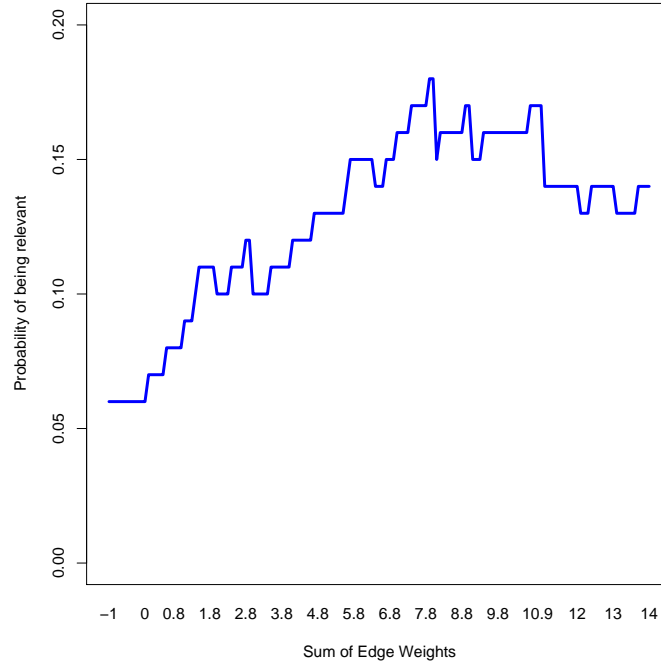


FIGURE 4.7: Probability of a document being relevant given a certain level of interaction. The y-axis represents probability that the video is relevant and the x-axis represents assigned interaction value in our graph

The relevance value on the x-axis thus represents the sum of the weights of all of the edges leading to a particular node. The average interaction value was just 1.23, with non-relevant documents having an average value of 1.13 and relevant documents having an average of 2.94. This result is encouraging as it shows that relevant documents do receive more interaction from the users of the system. It can be seen that up until a certain point as the interactions from previous users increase so does the probability of the document being relevant. It was also found that for some of the documents with higher relevance values the probability tails off slightly. Further investigation found that there were two main reasons that a number of irrelevant documents had high relevance values. Firstly, there were shots that seemed relevant at first glance but upon further investigation were not relevant; however, for participants to investigate this required some interaction with the shot thus giving it a high interaction value. Secondly, there were a number of shots that appeared in the top of the most common queries before any recommendations were given, thus increasing the chances of participants interacting with those videos. It should also be noted that on average only 5.49% of nodes in the graph relate to relevant shots. This indicates that users are exploring and interacting with large portions of the collection that are not relevant, to help them find relevant shots. However, even with this kind of sparse and difficult data the performance of the users is improved with the recommendations presented to the users. It was found that as the amount of information in the graph increased so did the proportion of recommendations selected by users; users 5–8 selected 9.77% of the recommendations, whereas users 21–24 selected 18.67% of the recommendations.

Text Queries

In both the baseline and recommendation systems the participants were presented with query expansion terms that they could use to enhance their queries. We found however, that the majority of participants chose not to use the query expansion terms provided by the baseline system as they found them confusing. The query terms returned by the baseline system were stemmed and normalised and hence were not in the written form as users expected them to be, where as the queries recommended by the recommendation system were queries that previous users had used. One participant stated that “The query expansion terms did not have any meaning.” Another participant said that the “query expansion did not focus on real search task”. This can be explained in part by specificities of some of the chosen topics, for example, in Task 1, when a user enters the name of a city (“New York”) to get a shot of the city’s sky line, the query expansion terms did not help to specify the search query. The top five queries for each task are

4.5. Results

presented in Table 4.4.

TABLE 4.4: Five most popular queries for each task

Task 1		Task 2		Task 3		Task 4	
City	9	Jail	5	Flag	8	Kiss	22
New York	8	prisoner guards	4	meeting flag	7	greeting kiss	20
tall buildings	8	prisoner	4	conference	5	greeting	10
Tower	7	prisoner escorted	3	group flag	5	cheek	6

Across all 24 users a number of terms were repeated extensively. There were 130 unique terms and combinations of these were used to create a number of unique queries, on average the participants used 2.21 terms per query. However, it can be seen that across the 24 users and 4 topics there is relatively little repetition of the exact same queries, there were 621 unique queries out of 1,083 total queries (57%). In fact only 4 queries occur 10 times or more, and they were all for the same task. This task had fewer facets to it than the others, and thus there was less scope for the users to use different search terms. This shows that despite the fact that users are carrying out the same task they are searching in differing ways, as the search tasks are multi-faceted and the participants are providing their own context. The results in this section indicate that the users explore the collection to a greater extent using the recommendations. Nodes were added to the graph of implicit actions throughout the evaluation (see Table 4.3). Also there was very little query repetition, and newer users used new and diverse query terms. These results give an indication that further participants are not just using the recommendations to mark relevant videos, but also interacting with further shots. These results also give an indication that Hypothesis H_4 holds; that users will be able to explore the collection to a greater extent, and also discover aspects of the topic that they may not have considered. However, this finding has not been fully validated. In order to do this we must analyse the users' perceptions of the tasks, this analysis is presented in the following section.

4.5.3 User Perception

In order to provide further validation for Hypothesis H_4 that “the users will be able to explore the collection to a greater extent, and also discover aspects of the topic that they may not have considered”, and to validate Hypothesis H_5 , that “the users will be more satisfied with the system that provides feedback, and also be more satisfied with the results of their search”, we analysed the post task and post experiment questionnaires that our participants filled out.

Search Tasks

To begin with, we wished to gain insight into the participants' perceptions of the two systems and also of the tasks that they had carried out. In the post-search questionnaires, we asked subjects to complete four 5-point semantic differentials indicating their responses to the attitude statement: "The search we asked you to perform was". The paired stimuli offered as responses were: "relaxing"/"stressful", "interesting"/"boring", "restful"/"tiring" and "easy"/"difficult". Using these differentials, we aimed to cover different possible semantics that the participants might use to describe their own perception of the search process. Similar differentials can be found in the research of White [2004] and Urban [2007]. The average obtained differential values are shown in Table 4.5 for each system.

TABLE 4.5: Perceptions of search process for each system (higher = better)

Differential	Baseline	Recommendation
Easy	1.9	2.65
Restful	2.7	2.575
Relaxing	2.725	3.175
Interesting	2.325	2.75

Each cell in Table 4.5 represents the responses for 20 participants (the four participants in the initial training set were not included as they did not use the recommendation system). The most positive response across all system and task combinations for each differential pair is shown in bold. The trends in Table 4.5 indicate that the users gave more positive responses for the recommendation system. It was found that the participants perceived some tasks as more easy, relaxing, restful and interesting than others. It can also be seen in Table 4.5 that there is a slight preference toward the system that provides recommendations amongst the participants. We applied two-way analysis of variance (ANOVA) to each differential across both systems and the four tasks. We found that how easy and relaxing the participants found the tasks was system dependent, whereas the user interest in the task was more dependent on the task that they were carrying out ($p < 0.194$ for the significance of the system).

Retrieved Videos

In post search task questionnaires we also solicited subjects' opinions on the videos that were returned by the system. We wanted to discover if participants explored the video collection more based on the recommendations or if it in fact narrowed the fo-

4.5. Results

cus in achievement of their tasks. The following Likert 5-point scales and semantic differentials were used:

1. “During the search I have discovered more aspects of the topic than initially anticipated” (Change 1)
2. “The video(s) I chose in the end match what I had in mind before starting the search” (Change 2)
3. “My idea of what videos and terms were relevant changed throughout the task” (Change 3)
4. “I believe I have seen all possible videos that satisfy my requirement” (Breadth)
5. “I am satisfied with my search results” (Satisfaction)
6. Semantic differentials : The videos I have received through the searches were: “relevant”/“irrelevant”, “appropriate”/“inappropriate”, “complete”/“incomplete”, “surprising”/“expected”.

Table 4.6 shows the average responses for each of these scales and differentials, using the labels after each of the Likert scales in the list above, for each system.

TABLE 4.6: Perceptions of the retrieval task for each system (higher = better)

Differential	Baseline	Recommendation
Change 1	3.1	3.5
Change 2	3.475	3.725
Change 3	2.725	3.05
Breadth	2.625	3.075
Satisfaction	2.95	3.4
Relevant	1.925	2.55
Appropriate	3.125	3.775
Complete	2.225	2.5
Surprising	1.55	1.725

The values for the four semantic differentials are included at the bottom of the table. The most positive response across all system and task combinations is shown in bold. The general trends that can be seen in Table 4.6 show that the users gave more positive responses for the recommendation system. It appears that participants have a better perception of the video shots that they found during their tasks using the recommendation system. It also appears that the participants believe more strongly that this system

changed their perception of the task and presented them with more options. This would back up the findings in the previous section that the participants explored the collection to a greater extent when presented with the recommendations. We applied two-way analysis of variance (ANOVA) to each differential across both systems and the four tasks to test these assertions. The initial ideas that the participants had about relevant shots were dependent on the task. The changes in their perceptions were more dependent on the system that they used rather than the task, as was the participants belief that they had found relevant shots through the searches. This demonstrates that the recommendation system helped the users to explore the collection to a greater extent, and also indicates that the users have a preference for the recommendation system. This finding strengthens the argument that the recommendation model is providing benefits in terms of exploration and user perception.

System Support

We also wanted to determine the participants' opinion about how the system supported their retrieval tasks. Firstly we asked them if the system had helped them to complete their task (satisfied). Participants were then asked to complete a further five 5-point Likert scales indicating their responses to the following statement: "The system helped me to complete my task because...". The criteria of the responses were:

1. "explore the collection" (explore)
2. "find relevant videos" (relevant)
3. "detect and express different aspects of the task" (different)
4. "focus my search" (focus)
5. "find videos that I would not have otherwise considered" (consider)

Table 4.7 presents the average responses for each of these scales, using the labels after each of the Likert scales above for each system. The most positive response is shown in bold. Some of the scales were inverted to reduce bias in the questionnaires.

Once again it appears that participants have a better perception of the video shots that they found during their tasks using the system with recommendations, and that they believe the system helped them to explore the collection of shots more thoroughly using this system. We applied two-way analysis of variance (ANOVA) to each differential across both systems and the four tasks to test our hypotheses; none of the dependencies were significant. From our analysis of the results, however, there is a trend that the

4.5. Results

TABLE 4.7: Perceptions of the system support for each system (higher = better)

Differential	Baseline	Recommendation
Satisfied	3.3	3.6
Explore	3.775	3.9
Relevant	3	3.4
Different	2.925	3.275
Focus	2.625	3.25
Consider	3.075	3.375

focus of the search, the ability to express different aspects of the task and the change in videos considered is more dependent on the task, rather than the system.

Ranking of Systems

After completing all of the tasks and having used both systems we attempted to discover whether the participants preferred the system that provided recommendations or the system that did not. The participants were asked to complete an exit questionnaire where they were asked which system they preferred for particular aspects of the task, they could also indicate if they found no difference between the systems. The participants were asked, “Which of the systems did you...”:

1. “find best overall” (best)
2. “find easier to learn to use” (learn)
3. “find easier to use” (easier)
4. “prefer” (prefer)
5. “find changed your perception of the task” (perception)
6. “find more effective for the tasks you performed” (effective)

The users were also given some space where they could provide any feedback on the system that they felt may be useful.

Table 4.8 shows that the participants had a preference for the system that provided the recommendations. It is also encouraging that the participants found there to be no major difference in the effort and time required to learn how to use the recommendations that are provided by the system with recommendations. This indicates that users were more satisfied with the system that provides recommendation, thus supporting Hypothesis H_5 . Users have a definite preference for the recommendation system. 17 out

4.6. Summary

TABLE 4.8: User preferences for the two different systems

Differential	Baseline	Recommendation	No difference
Best	2	16	1
Learn	2	7	11
Easier	2	5	13
Prefer	1	17	2
Perception	3	11	6
Effective	3	14	3

of 20 users preferred the recommendation system, with one user preferring the baseline system. The participants also indicated in their post task questionnaires that the system that provided recommendations helped them to explore the task and find aspects of the task that they otherwise would not have considered, in comparison with the baseline system. Since both hypotheses Hypothesis H_4 and Hypothesis H_5 are supported, we further conclude that Hypothesis H_3 can be supported. Implicit relevance feedback has led to the recommendation of relevant video documents that allowed the users to explore the collection further and helped them to satisfy their information need. The results of our analysis have addressed all of the points of our hypotheses and have demonstrated that we have achieved our goals.

4.6 Summary

In order to support the results of the performed user simulation that have been presented in the preceding chapter, we introduced in this chapter a user study where implicit relevance feedback is used to recommend video shots. The chapter has focused on validating various hypotheses. We have presented a novel approach for combining implicit and explicit feedback from previous users to inform and help users of a video search system. The recommendations provided are based on user actions and on the previous interaction pool. This approach has been realised in a video retrieval system, and this system was tested by a pool of users on complex and difficult search tasks. Using this setting, we aimed to model search scenarios where multiple users individually search for similar topics, as it is possible when using Web applications such as YouTube or Dailymotion.

For the results of task performance, whether users retrieve more videos that match their search task, we measured P@N and MAP values. It was shown that the recommendation system outperforms the baseline system, in that the users of the recommendation system retrieve more accurate results overall and that this difference is statistically sig-

4.6. Summary

nificant. It was also seen that users are finding relevant results more quickly using the recommendation system. These results validate our Hypothesis H_4 , since the performance of users of the recommendation system improved with the use of recommendations based on implicit feedback. The statistics presented on user exploration, show that the users are pursuing the tasks sufficiently differently. They were able to explore the collection to a greater extent and find more relevant videos. Nodes were added to the graph of implicit actions throughout the evaluation, indicating that users are not just using the same queries and marking the same results, but they are exploring new parts of the collection. These results give an indication that further participants are not just using the recommendations to mark relevant videos, but also interacting with further shots. This also supports Hypothesis H_4 ; that users will be able to explore the collection to a greater extent, and also discover aspects of the topic that they may not have considered. In addition to demonstrating the validity of this hypothesis, these findings also illustrate the validity of our approach and experimental methodology. The tasks that were chosen for the experiment were multi-faceted and ambiguous. As the tasks are multi-faceted we believed that participants would carry out their searches in differing ways and use numbers of different query terms and methodologies, thus providing their own context. This belief has been demonstrated by these findings. Hypothesis H_4 was validated by our analysis of user perceptions of the system where the users gave an indication that the recommendation system helped them to explore the collection. The participants indicated in their post task questionnaires that the system that provided recommendations helped them to explore the task and find aspects of the task that they otherwise would not have considered, in comparison with the baseline system. It is also shown that the users have a definite preference for the recommendation system. 17 out of 20 users preferred the recommendation system, while one user preferred the baseline system. These findings support Hypothesis H_5 that users will be more satisfied with the system that provides feedback, and also be more satisfied with the results of their search. These results successfully demonstrate the potential of using this implicit feedback to aid multimedia search, supporting Hypothesis H_4 that “implicit relevance feedback can be employed to recommend relevant video documents.”

Whilst the results indicate that implicit relevance feedback can effectively be used in the context of short term user profiling, various limitations of this study need to be highlighted. First of all, the experimental methodology is artificial: Users were asked to retrieve as many documents as possible to a pre-defined search task. Even though this is the de-facto standard evaluation methodology in interactive IR, it is questionable whether users will interact in the same way when following their own agenda. A detailed discussion on this problem is given in Section 2.3.3. Secondly, users often do

4.6. Summary

not perform single search sessions, but often perform searches over multiple iterations. In the next chapter will therefore consider whether implicit relevance feedback can also be employed to track user interest over a longer period of time.

– *In the future personalised search will be one of the traits of leading search engines.*

Marissa Mayer, 2008

5

Capturing Long-Term User Information Needs

Chapters 3 and 4 suggest that implicit relevance feedback can effectively be exploited to recommend relevant video documents. This chapter explores how such feedback can be used for implicit user profiling. After giving an introduction in Section 5.1, we discuss the need for user profiling in Section 5.2. Section 5.3 discusses requirements that allow multimedia recommendation. In Section 5.4, we introduce the methodology for creating and exploiting user profiles. The chapter concludes with a discussion of this methodology in Section 5.5.

5.1 Introduction

The results of the previous chapter suggest that implicit relevance feedback can be applied in the video retrieval domain. Similar to text-based retrieval approaches, video retrieval systems can be built that personalise retrieval results based on user interactions. The advantage is that recommendations can be provided without forcing the users to explicitly provide feedback, which takes away a huge burden from the user. An example system has been introduced in Chapter 4, where implicit relevance feedback from multiple users has been exploited to generate video recommendations that match the user's current information need. The evaluation has shown that users' implicit relevance feedback can be used to improve both retrieval performance and user satisfaction

within one search session where users search for a single topic.

A survey conducted by [Morris et al. \[2008\]](#), however, reveals that such scenario is not very representative to real life search situations. Their survey illustrates that a vast majority of participants often performs search tasks spanning more than one session. Infact, 73% of all respondents of their survey reported that they regularly perform multi-session tasks which can be distributed over several days. In order to assist users within these multiple sessions, it is therefore necessary to keep track of their feedback, i.e. by creating a user profile. [Mustafa \[2005\]](#) argues that:

“New search engines are improving the quality of results by delving deeper into the storehouse of materials available online, by sorting and presenting those results better, and by tracking your long-term interests so that they can refine their handling of new information requests. In the future, search engines will broaden content horizons as well, doing more than simply processing keyword queries typed into a text box.”

In this chapter, we discuss terms and conditions that are required for the creation of such multi-session user profiles that capture users’ long-term interests. Further, we propose novel methods to approach these conditions. We first introduce in [Section 5.2](#) a fictitious application scenario in which a personalisation system automatically provides multimedia content that matches a user’s interest. The analysis of the scenario leads to the first contribution of this chapter, which is a framework that could be used to implement a system supporting this scenario. In [Section 5.3](#), we define the requirements that allow us to focus on studying the use of implicit relevance feedback for the generation of implicit user profiles. The main contribution is an approach to analyse multimedia content that eases user profiling and corresponding recommendation of multimedia content. [Section 5.4](#) introduces novel methodologies to tackle various research challenges towards the creation of implicit user profiles. We discuss the main challenge how can implicit relevance feedback techniques be exploited to create efficient profiles and, further, how such profiles should be structured to separate different long-term interests. The introduced techniques are evaluated in the subsequent chapters of this thesis.

In summary, we address the following research challenges and questions in this chapter:

- Q₁: What infrastructure is required to provide multi-session video recommendations?
- Q₂: How should the video corpus be prepared to allow multimedia recommendation?
- Q₃: How should a user profile be structured to effectively represent the user’s current interests?

The research which is presented in this chapter has been published in [Hopfgartner and Jose, 2010b; Elliott et al., 2009; Hopfgartner and Jose, 2009a,b; Hopfgartner et al., 2008b].

5.2 The Need for User Profiles

As discussed in Section 2.2, personalised retrieval systems exploit individual user profiles to adapt retrieval results or to recommend documents that match the user's information need. In this section, we illustrate various research challenges that apply in the generation of personalised multimedia retrieval systems. We first introduce a fictitious personalisation scenario, envisioned by Sebe and Tian [2007], in Section 5.2.1. The scenario provides a vivid introduction into challenges and research opportunities in the domain. After analysing the technical requirements of such scenario, we propose in Section 5.2.2 a framework for standardised long-term user modelling. The framework supports the collection of the stream of data generated by people during their daily activities and uses the evolving set of concepts in domain-specific ontologies to extract relationships between the different data produced by the different mediums. These raw data and the linking concepts between them can be exploited by collaborative filtering and content-based recommendation algorithms to help individuals receive information pertinent to their on-going daily activities. In Section 5.2.3, we discuss research challenges arising from long-term user modelling approaches and introduce limitations of the introduced application scenario.

5.2.1 Long-Term Personalisation Scenario

In recent years, various application scenarios have been introduced to frame research activities in the field of personalised multimedia retrieval, e.g. within the European projects EU-MESH [MESH, 2006], PHAROS [Paiu et al., 2008] and PetaMedia NoE [Lagendijk et al., 2009]. It shows the increasing attention within the research community towards personalised multimedia retrieval. In this section, we outline an example multimedia personalisation scenario, introduced by Sebe and Tian [2007], that emphasises arising challenges in the research field.

“John Citizen lives in Brussels, holds a degree in economics, and works for a multinational company dealing with oil imports. He enjoys travel with emphasis on warm Mediterranean sites with good swimming and fishing. When watching TV his primary interest is international politics, particularly European. During a recent armed conflict he wanted to understand

different perspectives on the war, including both relevant historical material as well as future projections from commentators. When he returns home from work, a personalized interactive multimedia program is ready for him, created automatically from various multimedia segments taken from diverse sources including multimedia news feeds, digital libraries, and collected analyst commentaries. The program includes different perspectives on the events, discussions, and analysis appropriate for a university graduate. The video program is production quality, including segment transitions and music. Sections of the program allow him to interactively explore analyses of particular relevance to him, namely the impact of war on oil prices in various countries (his business interest), and its potential affect on tourism and accommodation prices across the Mediterranean next summer. Some presentations may be synchronized with a map display which may be accessed interactively. John's behavior and dialogue with the display are logged along with a record of the information presented to allow the system to better accumulate his state of knowledge and discern his interests in order to better serve him in the future. When John is away from home for business or leisure, he may receive the same personalized information on his mobile device as well, emphasizing information reflecting the neighborhood of his current Mediterranean location."

In this scenario, a recommender system collates multimedia fragments from different sources to generate an interactive multimedia package which is tailored to a user's interests. It is not specified in the scenario how the recommender system determines the user's personal interests. However, the scenario's protagonist John Citizen might use a vast amount of desktop applications, web applications, and computing devices that are capable of capturing his personal interests and activities and thus could be used as input devices to generate a personalised user profile. For example, John's mobile phone can be used to identify his current location, e.g. via an integrated GPS chip or based on the network triangulation. Indeed, services such as Google Latitude⁵⁻¹ keep track of this information. Another device could be John's interactive TV box, which he uses to view "broadcasts on Mediterranean sites with good swimming and fishing". By allowing the system to constantly keep track of his television viewing habits, John provides implicit relevance feedback on his personal interests over multiple sessions. Services such as TiVo⁵⁻² exploit this television consumption behaviour to recommend similar television

⁵⁻¹<http://google.com/latitude/>, last time accessed on: 8 May 2010

⁵⁻²<http://www.tivo.com/>, last time accessed on: 8 May 2010

content [Ali and van Stam, 2004]. If the heterogeneous data streams that are capturing aspects of John's life could be collected into a single stream, he may be able to benefit from relationships between the different aspects. For example, if John travels to Greece, a recommendation system exploiting both information, his current location and his general interests in "Mediterranean sites with good swimming and fishing" could recommend magnificent local beaches and fishing grounds.

Harnessing these individual streams of life data represents an interesting research problem in the area of long-term user modelling and the exploitation of the data captured in these models. Some existing approaches for capturing such streams include MyLifeBits⁵⁻³ and friendfeed⁵⁻⁴. MyLifeBits is a research project into a lifetime story of everything inspired by the Memex personal information system. MyLifeBits stores content and metadata for many different types of data and describes them using multiple overlapping common attributes. Friendfeed is a web-service which collects data using the RSS or Atom publishing protocol standards from blogging services, bookmarking services, photo storage services, status update services, music services, and many other life data collection platforms. While these projects provide solutions on how to collate data from different sources, they do not outline how these different information streams can be exploited to provide personalised information. Thus, as one contribution of this chapter, we outline a generic framework that could be used as a guideline to implement systems supporting the scenario in the following section.

5.2.2 A Generic Framework for User Profiling

The previous section introduced an application scenario where multimedia content is adapted to a user's personal interests. In this section, we analyse the scenario further. Thus, the contribution of this section is a generic framework that allows the above scenario. It has been published as a position paper by Elliott et al. [2009]. Figure 5.1 shows a sketch of a potential multimedia personalisation system framework. It consists of four main components: a Long-Term User Model (LTUM) API, a Life-Log repository, a modelling component, and a recommendation component. The following paragraphs describe each component in detail.

LTUM API. The architecture of the Long-Term User Model (LTUM) provides an interface to support inputs of continuous life data streams obtained from an individual's devices. Further, it outputs recommendations presented in an appropriate manner with

⁵⁻³<http://research.microsoft.com/en-us/projects/mylifebits/>, last time accessed on: 8 May 2010

⁵⁻⁴<http://www.friendfeed.com/>, last time accessed on: 8 May 2010

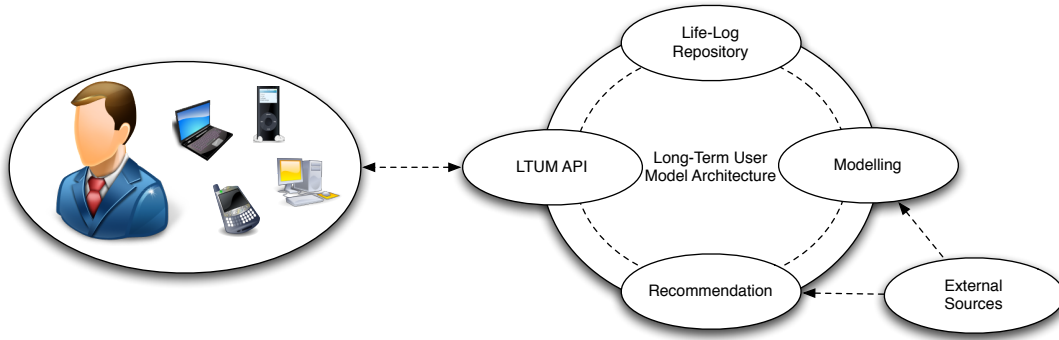


FIGURE 5.1: High-level architecture of life-long user modelling

respect to temporal and contextual awareness. This interface, referred to as the LTUM Application Programming Interface, allows our architecture to support many applications and devices. The LTUM architecture can collect the data generated by users in daily life activities and planned events in order to model the life data into a user profile with the surrounding context. Every device John uses would hence communicate with his user profile by using this API.

Life-Log Repository. Data captured from the incoming data stream reflects different aspects of users' personal interests. Aiming to support the example scenario, we propose a personal repository which can be automatically updated using any device. It would hence contain personal information such as GPS coordinates, music the user listened to, or feedback provided on recommendations. One challenge for this component is to filter important from unimportant information in the repository, for example by identifying events.

Modelling. Jain [2008] suggests modelling the user activities into distinct events, which can be defined by temporal, spatial, experiential, casual, informational and structural aspects. For example, an event could be that John enjoys listening to classical music on his way to work using his portable music player. Recently, the use of ontologies has been proposed to combine above aspects to related events. Pursuing this idea, we propose to use ontologies to exploit the repository accordingly. An important key for modelling of the user interests can be the use of external sources. If John, for instance, listens to audio tracks composed by *Haydn* and *Beethoven*, a user model can be enriched with the information that both are classical composers. Another challenge is to keep these events in a user profile. We propose to store such profiles on online profile servers that can interact with other people's profiles and/or external sources.

Recommendation. Once a user's interests have been identified and modelled accordingly, this knowledge can be used to recommend other information that might be of a user's interest. For example, if John's profile reveals a general interest in classical music and his current location, location-sensitive recommendations about classical performances occurring during John's stay in Greece could be suggested. Knowledge of the location of friends sharing a similar taste in music can lead to recommending that both could attend the concert together. This inter-linking, however, requires a careful investigation with respect to privacy issues since you might not always want to tell others where you are right now and what your interests are.

5.2.3 Discussion

The main contribution of this section was the definition of an infrastructure that supports multi-session video recommendations. Therefore, we discussed an application scenario where a multimedia recommender system generates personalised information to satisfy a user's information need. We further proposed a system framework that allows the implementation of recommender systems supporting the introduced scenario. As mentioned before, the application scenario is still a vision rather than a real-life scenario. Even though parts of the proposed framework are already in use, the whole technology is still in its infancy, since various research challenges need to be solved first. In the remainder of this section, we will discuss the research challenges of the different modules of the framework.

The proposed framework provides an API for various applications and devices to capture user activities. This can lead to a permanent, personalised data stream which can be stored in an online repository. As we discussed, existing applications such as friendfeed already allow the aggregation of multiple personal data streams. Providing a standardised API is hence not a scientific question any longer, but rather a technical or even business oriented problem.

Given such a repository, the framework suggests to generate user models by analysing the content of the repository, e.g. by identifying personal events. As studies (e.g. [Shaw et al. \[2009\]](#); [Scherp et al. \[2009\]](#)) indicate, further research is still required for a positive identification of events. The main problem is that insufficient information may be available to positively identify such events, or the information available may be vague or ambiguous. Even though it is now possible to handle ontologies that model large scale datasets consisting of a billion relationships between concepts [[Schenk et al., 2008](#)], an important challenge is still to find the correct representation for each concept in a user profile. Since our main research focus is on exploiting implicit relevance feedback in

the news video domain, we will not pursue this problem further.

A more relevant challenge for us is to analyse the multimedia content itself. A thorough analysis of the content users interact with is essential for the generation of personalised recommendations. The application scenario focuses on news about “a recent armed conflict”. We will therefore focus on daily news broadcasts, hence ignoring any other multimedia source that could be used for the personalisation of content. In Section 5.3, we discuss issues and problems arising when analysing multimedia content.

Another challenge is to identify those factors in the profile which should be employed for recommendation. The proposed architecture allows the use of internal and external sources, including friends’ profiles, to recommend information of interest. Considering these additional sources, privacy becomes a serious issue. It needs to be clarified to which degree each individual is willing to provide information that can be used by others. In an e-shopping scenario, Gálvez Cruz [2009] suggest to grant customers full control over their personal data when shopping online. Transferring her proposal to our scenario, users should be able to control, modify, and delete all details that are captured and used in a long-term profiling system. Note that we will not pursue these problems, since it is out of scope of this research.

Having the users’ interests captured in a profile, a challenge is how to identify this interest from their profile. One challenge is that users can show interest in multiple news topics. John Citizen, for example, is interested in European politics and Mediterranean countries. He further could be interested in sub categories such as Greek islands, Spanish beaches or Italian dolce vita. A specification for a user profile should therefore be able to automatically identify these multiple aspects. In Section 5.4, we introduce our methodology of multi-session user profiling and multimedia recommendation.

5.3 Requirements for a User Profile

John Citizen, the character in the application scenario introduced in the previous section uses a news recommender system that automatically generates personalised multimedia packages that cover topics of his interest. In the scenario, these packages consist of many different multimedia segments such as news feeds or commentaries. As we discussed, however, we simplify the example by considering news broadcasts only. When the packages are generated from up-to-date multimedia news broadcasts only, it comes clear that they can either be collections of relevant *shots* from a given news story or collections of relevant news *stories*. Shots are often used to visually enrich the actual news story, e.g. by providing impressions of the location of the event. Sometimes, even

archived video footage is used that is not in direct connection to the actual news. News stories consist of a series of shots. It is up to the editor of the television broadcast to decide which shots are used to report the news story. They should be seen as a means to assist the news consumer in understanding the news, rather than being the main unit conveying the news. We therefore define that the news video recommender system should focus on generating personalised multimedia packages consisting of news *stories*. We discuss in Section 5.3.1 challenges arising when focusing on news stories.

Another requirement for recommending news stories is to analyse the content of these news stories. This is, due to the Semantic Gap, not trivial though. As discussed in Section 2.1.2, a promising approach to ease this problem is to set such multimedia documents into their semantic contexts. For instance, a video about David Cameron's visit to Italy can be put into different contexts. First of all, it shows an event which happened in a Mediterranean country, Italy. Moreover, it is a visit by a European politician, the prime minister of the United Kingdom. If the fictitious John Citizen follows news about Cameron's visit, it might indicate that he is interested in either politics, Italy, or in both. Knowing the context of a video is useful for recommending other videos that match the users' information need. By exploiting these contexts, multimedia documents can also be linked to other, contextually related documents. Due to recent improvements in Semantic Web technologies, it is now feasible to automatically link concepts to the Linked Open Data Cloud, where they are connected to other concepts. Section 5.3.2 discusses this technology further. Any news story's concepts can hence be set into its semantic context. Based on the state-of-the-art research, we hypothesise that exploiting this context can lead to appropriate news video recommendations. Thus, the main contribution of this section is a novel methodology to set multimedia documents into their semantic context. We propose an approach of generating this semantic link in Section 5.3.2. In Section 5.3.3, we propose to categorise news stories based on their subject to ease access to the collection. Section 5.3.4 summarises and discusses the proposed requirements.

5.3.1 Capturing and Segmenting News Broadcasts

The most essential requirement for the previously presented multimedia recommendation scenario is to acquire up-to-date news broadcasts. In most countries, daily television news bulletins can be received by either aerial antenna or satellite dish. Recently, some television stations started to offer their news bulletins as online download, e.g. the *BBC One O'Clock News* on the BBC's iPlayer portal⁵⁻⁵ or the German *Tagesschau* as

⁵⁻⁵<http://www.bbc.co.uk/iplayer/>, last time accessed on: 8 May 2010

Podcast in the ARD Mediathek⁵⁻⁶. Consequently, a large amount of potential sources are available which could be used to create a personalised news broadcasting collection. The UK legislation on copyright and related rights⁵⁻⁷ allows us for single copies of copyrighted works for research or private study. Note that different copyright laws apply in each country. Cole [2009] discusses issues related to copyright in a digital context from a UK perspective.

The next step after capturing the daily broadcast is to automatically segment it into semantically coherent sequences. As discussed in Section 2.1.3, a consumer-oriented segmentation approach is to identify story boundaries. Note that news story segmentation is not the main focus of this thesis and will therefore not be discussed further here.

5.3.2 Exploiting External Knowledge

In Section 2.1.2, we discussed that multimedia documents are often enriched with additional metadata such as creation date, source or descriptive tags. The informative nature of news video broadcasts results in a large amount of potential textual tags, because News aim to provide a compressed overview of the latest events. Events are thus usually summarised by a background narrator, journalist or anchor person, resulting in text heavy transcripts. Indexing news videos based on such transcripts would enable textual retrieval and ease users' access to the corpus. Indeed, studies, e.g. within the evaluation workshop TRECVID, have shown that textual features are still the best source to perform multimedia retrieval [Christel, 2007a].

A closer analysis of state-the-art research within TRECVID, however, also indicates that the retrieval performance of news video retrieval is still far away from their textual counterparts. An interesting approach for narrowing this performance gap is to further enrich the textual transcripts using external data sources. Fernández et al. [2009], for instance, have shown that ontology-based search models can outperform classical information retrieval models at a Web scale. The advantage of these models is that external knowledge is used to place the content into its semantic context. Due to large community efforts such as the Linked Open Data project, broad collections of freely available concepts are available that are interlinked using different ontologies. As discussed, the backbone of this cloud is DBpedia, an information extraction framework which interlinks Wikipedia content with other databases on the Web such as Geonames or WordNet. Figure 5.2 illustrates that the DBpedia Knowledge Base is a graph of linked concepts. As of April 2010, it contains more than 2.6 million graph elements

⁵⁻⁶<http://www.ardmediathek.de/>, last time accessed on: 8 May 2020

⁵⁻⁷Copyright, Designs and Patents Act 1988, Chapter III section 29 – online available at: <http://www.ipso.gov.uk/cdpact1988.pdf>, last time accessed on: 15 March 2010

5.3. Requirements for a User Profile

which are interlinked with each other. The nodes are concepts that are identified by unique identifiers, URI's. A semantic hierarchy between most neighboured nodes is defined by the Simple Knowledge Organisation System Reference (SKOS) data model⁵⁻⁸. Figure 5.2 illustrates an example hierarchy, the hierarchy of the concept “Scotland” in DBpedia.

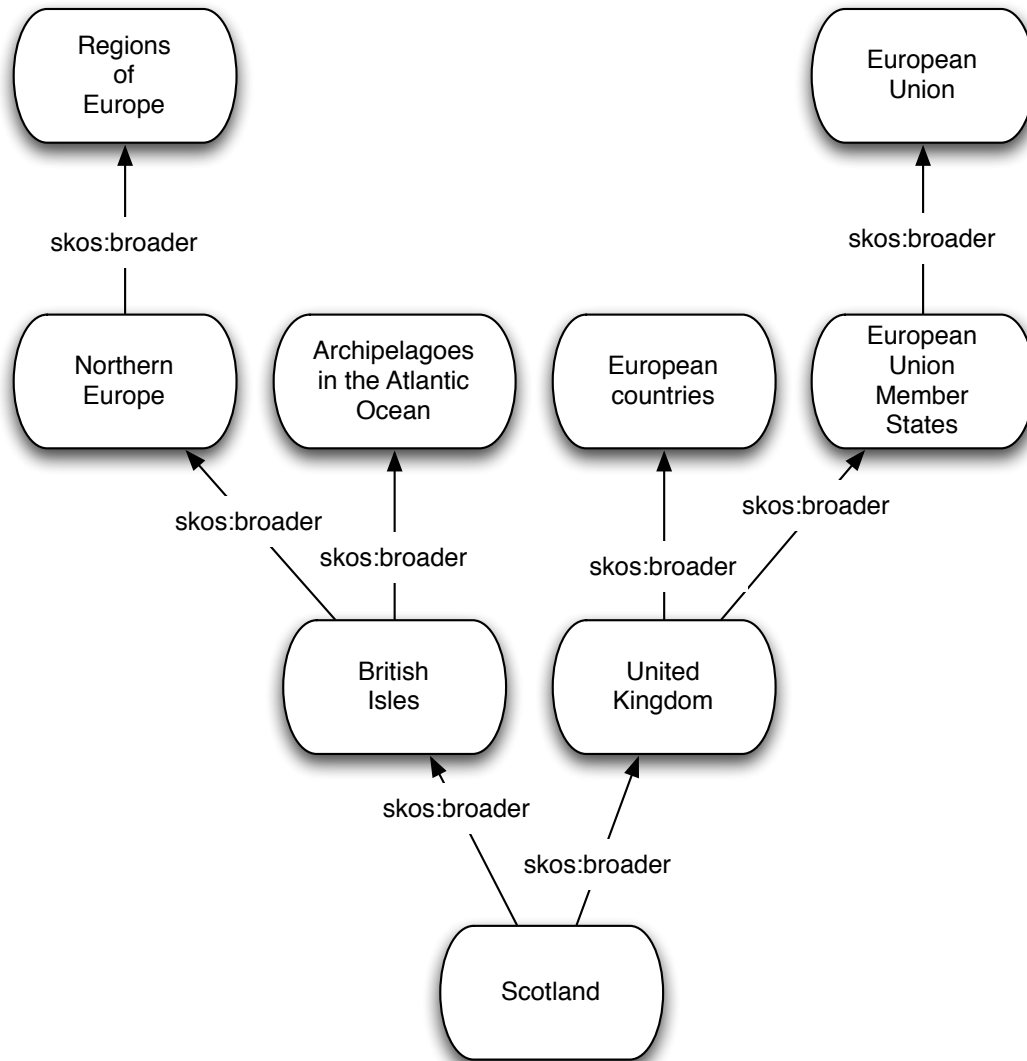


FIGURE 5.2: Hierarchy of the concept “Scotland” in DBpedia

From a news personalisation perspective, this semantic link provides the potential to improve interactive video retrieval and recommendation. For example, John Citizen could show interest in a story about the sunset at the Greek island Santorini. The story

⁵⁻⁸<http://www.w3.org/TR/skos-reference/>, last time accessed on: 8 May 2010

transcript might contain the following sentence:

“This is Peter Miller, reporting live from Santorini, Greece, where we are just about to witness one of the most magnificent sunsets of the decade. [...]”

Since John enjoys travel with emphasis on warm Mediterranean sites, he might also be interested in a report about the Spanish island Majorca. For example, imagine the following story:

“Just as every year, thousands of tourists enjoy their annual sun bath here in Majorca. [...]”

An interesting research question is how to identify whether this story matches John’s interests. Lioma and Ounis [2006] argue that the semantic meaning of a text is mostly expressed by nouns and foreign names, since they carry the highest content load. Indeed, as discussed in Section 2.2.4, most adaptation approaches rely on these terms to personalise retrieval results, e.g. by performing a simple query expansion. The two example stories, however, do not share similar terms. A personalisation technique exploiting the terms only would hence not be able to recommend John the second story.

However, as Figure 5.3 illustrates, linking the concepts of the transcripts using DBpedia reveals the semantic context of both stories. It becomes evident that both stories are about two islands in the Mediterranean Sea. Exploiting this link could hence satisfy John’s interest in warm Mediterranean Sites. We therefore propose to set news broadcasts into their semantic context by exploiting the large pool of linked concepts provided by DBpedia.

5.3.3 Categorising News Stories

Gans [2005] argues that modern times news reports can be categorised into various news categories such as Political news, Sports news or Entertainment news. For example, the following story transcript, taken from the BBC’s news broadcast of 3rd March, 2010, could be categorised as belonging to a “Entertainment News” category.

“Hollywood’s biggest night of the year is almost upon it. The 82th Oscars Ceremony is taking place in Los Angeles on sunday. Who is likely to walk away with the gongs? Will it be the box office hit Avatar or the gritty Iraqi hit The Hurt Locker. Rajesh Mirchandani is on the red carpet where preparations are under way. [...]”

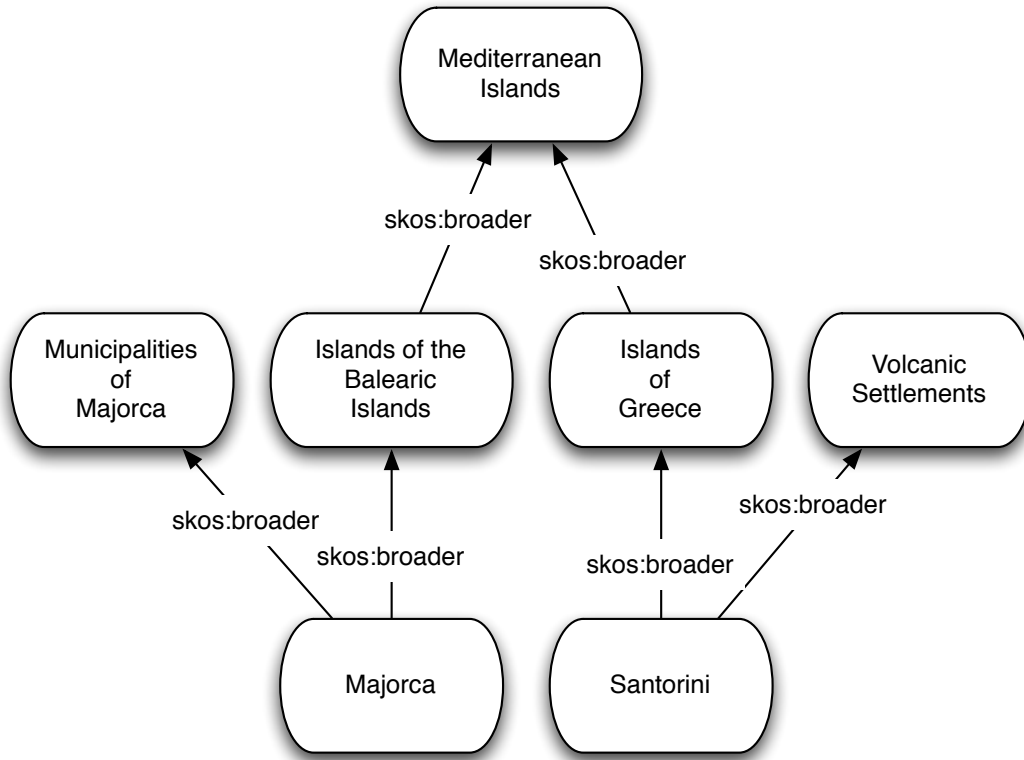


FIGURE 5.3: Linking “Santorini” and “Majorca” using DBpedia

Various approaches have been studied to automatically determine a news story’s subject and to categorise into such broad categories, e.g. [Joachims, 1998; Hayes et al., 1997; Diriyee et al., 2010]. The motivation for such a task is to ease access to the news corpus. Following this motivation, we suggest to categorise the news stories based on their subject since such categorisation could help to separate users’ interests.

5.3.4 Discussion

In this section, we discussed requirements that should be fulfilled to ease the generation of user profiles. We first discussed the creation of a private news video corpus, which requires capture and segmentation of daily news broadcasts. Further, we suggest the enrichment of resulting news stories by identifying and linking concepts using a generic ontology. Various problems need to be handled when using such an ontology. The main problem is how to automatically identify the correct concept for a given term. Shadbolt et al. [2006] argue in their survey on the development of semantic web technologies that this is the main problem within the domain and that in recent years, different techniques have been proposed. Since concept matching is out of scope for this research,

we will henceforth rely on these techniques, although this problem will be revisited in Chapter 6. Another challenge is the quality of the existing ontology. Being a representation of Wikipedia, both quantity and quality of DBpedia links differ tremendously. While some nodes have many neighbours, others are linked to only a few related concepts only. Further, the approach relies on the correctness of the information which is represented within DBpedia. Suchanek et al. [2007] manually evaluate the quality of DBpedia as part of their Yago ontology, a semantic knowledge base which builds on DBpedia. They report a fact correctness of 95%, suggesting that DBpedia can be seen as a reliable source. Another problem is that DBpedia is created automatically every six months from the English language version of Wikipedia. Its content is hence out of date, which might be a problem when news contain concepts that have not yet been described on Wikipedia or that were recently created only. Examples are public figures such as new politicians, successful business tycoons, new inventions or companies. Berners-Lee et al. [2001] envisioned the Semantic Web as consisting of machine readable information chunks which can be merged based on their semantic content. The current version of DBpedia can be seen as a milestone towards the development of such web. With the increasing success of the Semantic Web, chances increase that DBpedia (or similar approaches such as the Semantic MediaWiki project [Krötzsch et al., 2006]) can become an essential part of the Wikipedia project. A desirable improvement would then be to update DBpedia concepts in real time whenever a Wikipedia page has been changed. Such a technical advance would bridge the problem of an out-dated concept corpus. Nevertheless, major broadcasters such as the BBC [Kobilarov et al., 2009] now already rely on DBpedia, indicating that even an out dated corpus can serve successfully to link documents. Finally, we suggest to categorise news stories into broader news categories. These categories could ease the generation of user profiles.

5.4 Tackling User Profiling Problems

In the previous section, we suggested creation of a private news video collection consisting of up-to-date news bulletins from different broadcasting stations. Further, we introduced our approach of exploiting the Linked Open Data Cloud to link concepts in the news broadcasts and suggested a categorisation of stories into broad news categories. From a user profiling point of view, these links and categories can be of high value to recommend semantically related transcripts, hence creating a semantic-based user profile. In this section, we introduce our methodology of identifying user's long-term interests.

As explained in Section 2.2.2, providing feedback on a document is considered as

evidence that this document is relevant for the user's current interest. Most personalisation services rely on users explicitly specifying preferences. However, users tend not to provide constant explicit feedback on what they are interested in. In a long-term user profiling scenario, this lack of feedback is critical, since feedback is essential for the creation of such profiles. As shown in the previous chapter, implicit relevance feedback can be used to balance this missing feedback. We therefore argue that user profiles should be automatically created by capturing users' implicit interactions with the retrieval interface (see [Hopfgartner and Jose, 2010a,b, 2009a]). Hence, our hypothesis is that implicit relevance feedback techniques can efficiently be employed to create implicit user profiles. The contribution of this section is thus a novel approach to generate such profiles. We discuss the generation process in Section 5.4.1.

Another challenge is to capture users' evolving interests in implicit user profiles. What a user finds interesting on one day might be completely irrelevant on the next day. In order to model this behaviour, we therefore suggest in Section 5.4.2 to apply the Ostensive Model of Developing Information Needs. Further, we argue for the automatic identification of multiple aspects of users' interests.

In Section 5.4.3, we highlight the need to identify different aspects of interest and introduce our approach to solve this problem. Section 5.4.4 discusses the introduced work.

5.4.1 User Profile Model

The analysis of representative video retrieval interfaces in Chapter 3 revealed six user interactions which are commonly supported by most video retrieval interfaces. By treating these interactions as implicit indicators of relevance, we have shown that this implicit relevance feedback can be successfully employed to improve interactive video retrieval performed within single search sessions. The next challenge which needs to be addressed is how implicit relevance feedback techniques can be exploited to enable an application scenario as described in Section 5.2.1, that is the creation of efficient user profiles by implicitly capturing user interest. Unfortunately, the introduced scenario makes it almost impossible to capture their interests though. For example, the protagonist John might inform himself about the latest political developments in the newspaper, or by listening to the radio news. Consequently, he might not watch the news on television, since he is already aware of the current situation. Since we want to study whether implicit relevance feedback can be used to generate user profiles, we focus on the users' interactions with a news video recommender system, hence ignoring other sources.

By interacting with the graphical user interface of such system, users leave a "se-

mantic fingerprint” indicating their interest in the content of the items they have interacted with. As discussed in Section 3.4, the degree of their interest can be expressed by a weighting aligned with the different interface feature types. For example, the more interactions are performed on an item, the higher the weighting for this item, and the stronger the fingerprint that the user is interested in its content. The first challenge we then have to approach is how to capture this fingerprint.

In Section 2.2.2, we surveyed several user profiling approaches. A prominent way of capturing user interests is the weighted keyword vector approach. In this approach, the interests are represented as a vector of weighted terms where each dimension of the vector space represents a term aligned with a weighting. Considering the high semantics conveyed by each story users might interact with, generating user profiles on a term-based level only would ignore these semantics though. We therefore suggest a *weighted story vector approach* where each interaction I of a user i at iteration j of their search session is a vector of weights

$$\vec{I}_{ij} = \{SW_{ij1} \dots SW_{ijs}\}$$

where s indexes the story in the whole collection. The weighting SW of each story expresses the evidence that the content of this story matches the user’s interest. The higher the value of SW_{ijs} , the closer this match is.

Different from short-term adaptation services, a multi-session personalisation system requires the storage of the user’s semantic fingerprint. The next challenge is hence to store this vector of weights in a user profile. As explained in Section 5.3.3, we suggest classification of news stories into broad categories. This categorisation can be exploited to model the user’s multiple interests. For example, the character John from the previous scenario shows interests in both European Politics and Mediterranean countries. Having all interests in one profile is not effective; Since these are two different issues, it is reasonable to treat them separately. We therefore suggest to use this classification A as a splitting criteria. Thus, we represent user i ’s interest in an aspect A in a category profile vector $\vec{P}_i(A)$, containing the story weight $SW(A)$ of each story s of the collection:

$$\vec{P}_i(A) = \{SW(A)_{i1} \dots SW(A)_{is}\}$$

Each $\vec{P}_i(A)$ hence contains a vector of stories that belong to the aspect in category A and in which the user showed interest in at iteration j . This interest is expressed by the story weight SW , which is determined based on the implicit indicator of relevance which the user used to interact with the story.

Even though the user’s interests can be split into different broad categories, two

main problems remain. The first challenge is the capturing of user's evolving interest. Section 5.4.2 introduces our approach to handle this problem. The second challenge is the capturing of different sub aspects of this interest. This problem is tackled in Section 5.4.3.

5.4.2 Capturing Evolving Interest

The simplest approach to create a weighting for each story in the profile is to combine the weighting of the stories over all iterations. This approach is based on the assumption that the user's information interest is static, which is, however, not appropriate in a retrieval context. The users' information need can change within different retrieval sessions [Psarras and Jose, 2007, 2006; Elliott and Jose, 2009]. Following Stvilia et al. [2007]'s argumentation that information quality is sensitive to context changes such as time, an interesting research question is how this change of interest can be incorporated.

As discussed in Section 2.2.2, Campbell and van Rijsbergen [1996] propose to consider a time factor when ranking documents, i.e. by modifying the weighting of terms based on the iteration in which user's interacted with the corresponding document. They distinguish between four different functions to calculate the weighting, depending on the nature of aging:

- Constant weighting
- Exponential weighting
- Linear weighting
- Inverse exponential weighting

Considering the ostensive evidence as a method to model user interest in documents belonging to category A in a profile, we propose to manipulate the story weight in our category profile. Therefore, we define the story weight for each user i as the combination of the weighted stories s over different iterations j :

$$SW(A)_{is} = \sum_j a_j W_{ijs} \quad (5.1)$$

We include the ostensive evidence, denoted a_j , to introduce different weighting schemes based on the ostensive model. Figure 5.4 plots this evidence a_j for up to ten iterations. It can be seen that all functions, apart from the constant weighting, reduce the ostensive weighting of earlier iterations. The weighting depends on the constant $C > 1$.

In the remainder of this section, we discuss the functions in detail.

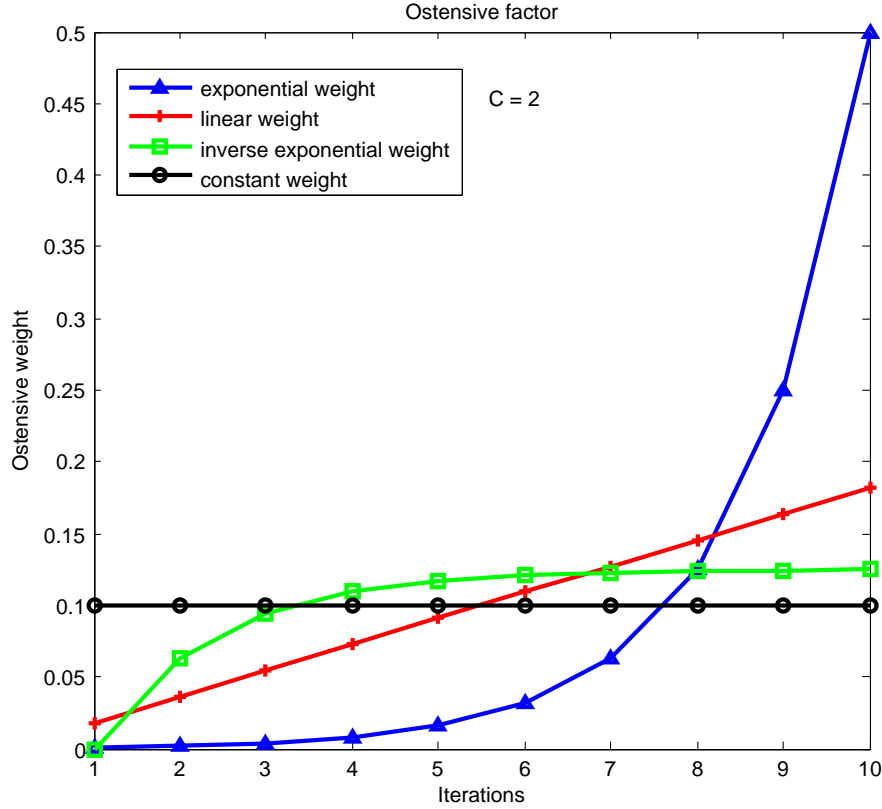


FIGURE 5.4: Effect of different ostensive weighting functions over ten iterations

Constant Weighting

$$a_j = \frac{1}{j_{max}} \quad (5.2)$$

The constant weighting function does not influence the ostensive weighting. As Equation 5.2 illustrates, all terms will be combined equally, ignoring the iteration when a term was added or updated. The constant weighting can be seen as a baseline methodology which does not include any ostensive factor.

Exponential Weighting

$$a_j = \frac{C^j}{\sum_{k=2}^{j_{max}} C^k} \quad (5.3)$$

The exponential weighting as defined in Equation 5.3 gives a higher ostensive weighting to terms which have been added or updated in older iterations. It is the most extreme function as the ostensive weighting of earlier iterations decreases steeply.

Linear Weighting

$$a_j = \frac{Cj}{\sum_{k=2}^{j_{max}} k} \quad (5.4)$$

Equation 5.4 defines the linear weighting function. The ostensive weighting of earlier iterations decreases linearly. This function linearly reduces the ostensive weighting of earlier iterations.

Inverse Exponential Weighting

$$a_j = \frac{1 - C^{-j+1}}{\sum_{k=2}^{j_{max}} (1 - C^{-k+1})} \quad (5.5)$$

The inverse exponential weighting defined by Equation 5.5 is the most discriminating function. Compared to the other introduced functions, the ostensive weighting of early iterations decreases more slowly.

Discussion

This section has shown how the different weighting functions of the Ostensive Model of Evolving Information Need could be applied in our multi-session user profile. As discussed, each function supports different usage scenarios. In the context of modelling users' interests in news, we can, however, reduce the number of functions that fit into such scenario. We assume that a user's interest in certain news evolve over time. Consider, for example, the following scenario. At the beginning of the credit crunch, John was following news about the financial troubles of big national banks. He showed some interest in it, but ignored the story after a while, since he was not directly affected. Day by day, however, more and more news appear, highlighting the complications of this bankruptcy case. At the same time, John's interest in the issue increases. Thus, his interest evolves from *low* interest to *high* interest.

A constant weighting does not provide any additional weighting to the user's profile. Exploiting corresponding profiles would hence not distinguish between recent and old feedback. The function thus can not be used to model an evolving interest. Likewise, a profile created using an inverse exponential weighting does not support such a scenario. Both exponential weighting and linear weighting, however, consider more recent feedback as stronger indicator of the user's interests than older feedback. While with the linear weight, this recent interest decays rather slowly over several iterations, the exponential weighting function decays very fast. Considering the nature of news, especially the sudden appearance of breaking news, the exponential weighting model seems to be the best function to model the user's evolving interest. Breaking news of rare events

such as an earthquake will not have any similar stories in the profile; nevertheless, they might be of high importance for the user. Consequently, their weighting should be considerably higher than the weighting of other news. The linear weighting function gives a relative high weighting to stories of earlier iterations. Thus, the breaking news story might perish. The exponential weighting function gives a relatively high weighting to more recent iterations. Breaking news would hence be ranked higher in the user profile. As argued before, [Campbell, 2000; Urban et al., 2006b; Leelanupab et al., 2009b] applied the Ostensive Model in the image domain, while Joho et al. [2007] introduce the model in a web scenario. In each case, the authors incorporate an exponential weighting as a decay function. Corollary, we propose to model this evolving interest in news by incorporating the exponential weighting as a decay function. An analysis of different approaches is given in Hopfgartner et al. [2008a].

5.4.3 Capturing Different Apects of Interest

In Section 5.4.1, we suggest to split user profiles into broad news categories which can be derived from the classification procedure suggested in Section 5.3.3. Whenever a user interacts with a news document, the category vector $\vec{P}_i(A)$ of the corresponding category is updated with the new story weight. As some stories might belong to more than one broad category, we propose to add the corresponding story to every associated category profile. For example, a news story about the Government’s decision to bail out domestic banks might be categorised as belonging to both categories Politics and Business. The drawback of this approach is that during the early stages of the user profiling process, semantically related category profiles might contain exactly the same news documents. Thus, at later stages of the user’s interactions, other documents might be added to either of the categories, resulting in different contents. These early duplicates are hence more a cold start problem that will decay in the process of user profiling.

The methodology introduced above results in a category-based representation of the user’s interests. Each category profile consists of a list of weighted stories, with the most important stories having the highest weighting. Following this approach, the profile of the example user John would consist of two main category profiles: Politics and Tourism. Since these are very broad news categories, however, each category profile might still be very diverse. News reports about European Politics, for instance, might be about the internal politics of different European countries or about different aspects such as the negotiation of new trading agreements or the installment of new immigration regulations. Likewise, stories in the user’s Tourism profile might contain stories about Greece or Spain or different activities such as fishing and swimming.

Aiming to exploit the user profiles to recommend the user other stories that are related to each of these sub categories presents a challenge of identifying different contextual sub aspects in their category profiles. This is an information filtering problem. Information filtering techniques exploit the fact that semantically related documents share certain textual features. Hence, clustering the documents based on these textual features should result in semantically related clusters where each cluster represents a sub categories of the user's interest. We therefore suggest the clustering of each category profile, and consider each cluster as a sub category.

5.4.4 Discussion

In this section, we introduced the methodology for generating implicit user profiles. Our main research focus was on discussing how implicit relevance feedback can be exploited for the generation of such profiles. We proposed to store news stories that the user interacted with, aligned with an implicit feedback weighting, in structured user profiles. Each time a user provides new implicit relevance feedback, the corresponding user profile is updated. It is an iteration-based representation of the user's interests. We further argued that this interest is not static, since users will lose interest in old news. Aiming to smooth this decay, we proposed to apply the Ostensive Model of Evolving Information Need. We discussed different ostensive weighting functions and argue for the use of a decay function. Another problem we addressed is the user's interests in multiple topics. We first argued for the generation of category-based user profiles to separate these interests. Categories can be derived from the news stories the users interacted with. Whenever a user interacts with a news story, the corresponding category profile is updated automatically. Since these categories might be very broad, we further argued to identify sub categories by clustering the content of each category profile. Each cluster then represents the user's interest in a certain aspect of the broad news category.

5.5 Summary

In this chapter, we studied whether implicit relevance feedback can be employed to generate implicit user profiles in the multimedia domain. We, therefore, first discussed a visionary application scenario of a multimedia recommender system which generates personalised multimedia documents that satisfy a user's information need. The first contribution of this chapter is a generic framework for implicit user profiling. As we demonstrated, the scenario cannot easily be achieved by applying current state-of-the-

art techniques. Thus, we limited the conditions of the scenario, which allows us to focus on studying the use of implicit relevance feedback. Outlining the requirements for implicit user profiles, we proposed to generate personalised news video collections and argued to process this collection by segmenting each bulletin into semantically coherent news stories, categorising them into broad news categories and by enriching these stories using a generic ontology. This data augmentation allows us to set news stories into their semantic context, which we then suggested to consider when creating the user profile. Thus, the second contribution of this chapter is a novel methodology to create such semantic link. Applying the results of Chapter 3, we suggested to use implicit relevance feedback to store relevant news stories in a profile. Further, we suggested to apply the Ostensive Model of Evolving Information Need to compensate the user's losing interest in stories he or she showed interest in during earlier stages. Finally, we discussed that the user profiles should be split based on the user's different interests and suggest to perform clustering to identify the user's interests.

An interesting question is how these clusters, that consist of documents that the user interacted with should be exploited to generate useful recommendations. In the next chapter, we will therefore evaluate different recommendation approaches which are based on the introduced user profiling methodology.

– *The world you perceive is a drastically simplified model of the real world.*

Herbert Simon, 1947

6

Simulation-based Evaluation of Long-Term News Video Recommendation Techniques

In the previous chapter, we outlined conditions for the generation of user profiles that capture users' long-term interests in news videos which they expressed implicitly over multiple search sessions. In this chapter, we evaluate how such profiles could be exploited to provide personalised news video recommendations that match the users information need. We therefore first illustrate our approach to fulfil the requirements for implicit user profiling in Section 6.2. In Section 6.3, we introduce different recommendation techniques. We aim to study these approaches by employing a simulation-based evaluation scheme which allows fine tuning various parameters. Such evaluation requires the generation of relevance assessment data which is discussed in Section 6.4 and the simulation of long term user profiles, which we discuss in Section 6.5. The simulation-based evaluation of the recommendation techniques is introduced and discussed in Section 6.6. Section 6.7 summarises and concludes this chapter.

6.1 Introduction

In Chapter 3, we showed that implicit relevance feedback, even when given by other people searching for similar content, can be exploited to recommend video shots that

satisfy the users' information need within one search session. We hypothesised, in Chapter 5, that not only can implicit relevance feedback be exploited to adapt retrieval results within one search session, but can also be employed to adapt users' interests over multiple sessions. In Section 5.2, we introduced an application scenario where a user benefits from a multimedia personalisation system which automatically generates multimedia programmes, consisting of news videos, that match his personal interest in news. Keeping track of users' interests over a longer time period requires storing their interests in a user profile. We therefore discussed various problems that need to be solved toward generating such profiles. First of all, we argued for the generation of a personal news video collection, consisting of the latest news broadcasts. We suggested categorisation and annotation of these news stories to ease handling its content. In Section 6.2, we introduce our approach to fulfil these requirements.

In Section 5.4, we proposed to capture users' interest in a profile. One challenge is that users can show interest in multiple news topics. For example, users may be interested in Sports and Politics or in Business news. Further, they can even be interested in sub categories such as Football, Baseball or Hockey. A specification for a implicit user profile should therefore be able to automatically identify these multiple aspects. We suggested to separate user profiles based on broader news categories. Moreover, we suggested that a hierarchical agglomerative clustering of the content of these category-based profiles can be used to effectively identify sub categories. As explained, the proposed user profiling approach gives a higher weighting to those stories that achieved a higher attention by the user. Two questions arise from this. The first question is, how many entries in such user profile should be used to represent the user's current topics of interest. Moreover, another question is how to exploit the identified sub categories of the profile in order to recommend relevant news stories that match the user's interest. In the previous chapter, we therefore suggested to automatically link concepts within the news stories to the Linked Open Data Cloud, where they are connected to other concepts. Any news story's concepts can hence be set into its semantic context. Based on the introduced related work, we hypothesise that exploiting this context can lead to appropriate news video recommendations. In Section 6.3, we introduce different recommendation techniques. An open question is, however, how many concepts should be considered to identify similar news stories to recommend to the user.

In order to evaluate the quality of the recommendations over a longer time period, a long-term user experiment is required where users are free to use the system to satisfy their personal information need. The constrictions of laboratory-based interactive experiments with pre-defined search tasks do not allow such scenario, since users will not be able to search for the content they are really interested in. Consequently, a general

list of assessed documents cannot be used, since the user decides what topic he/she is searching for. Moreover, the evaluation of different parameters requires a larger number of runs. A user-centric evaluation is therefore inadequate, since it would require many users to repeat the same steps various times. Tackling these problems, we suggest a novel evaluation methodology which is based on user simulation. In Sections 6.4 and 6.5, we introduce our approach of generating required data for such evaluation. In Section 6.6 we then introduce the actual simulation of the recommendation approaches. Section 6.7 concludes this chapter.

In summary, we address the following hypotheses in this chapter:

H₆: Implicit relevance feedback techniques can be exploited to create efficient user profiles.

H₇: Ontologies can be exploited to recommend relevant news documents.

Further, we address the following research questions:

Q₄: How many entries in a user profile should be used to represent the user's current topics of interest?

Q₅: How many concepts should be considered to identify similar news stories to recommend to the user?

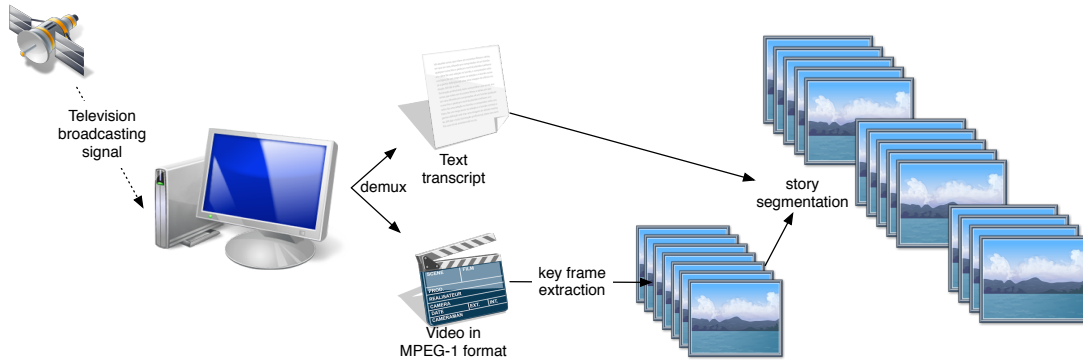
The research which is presented in this chapter has been published in [Hopfgartner and Jose, 2010b, 2009a, 2010a].

6.2 Generating a Daily News Corpus

In Section 5.3, we argued that a recommender system generating news packages should have access to an up-to-date news video corpus. We therefore suggested to create personal news video collections by capturing latest news bulletins. These bulletins need to be segmented into semantically coherent news stories. In Section 6.2.1, we introduce our approach for capturing the news stories and segmenting them accordingly. We further argued that these stories should be categorised based on broad news categories, since this eases recommendation approaches exploiting this data. In Section 6.2.2, we introduce our approach for determining these categories. In addition to categorising the corpus, we further proposed to set the news stories into their semantic concept by exploiting a generic ontology. We discuss our approach in Section 6.2.3. Section 6.2.4 summarises this section.

6.2.1 Capturing Daily News

FIGURE 6.1: News Capturing and Segmentation Process



Within this study, we focus on the daily BBC One O’Clock News and the ITV Evening News, the UK’s largest news programmes. Each bulletin has a running time of thirty minutes and is broadcast on work days. Both channels enrich their broadcasts with a closed caption (teletext) signal that provides textual transcripts. Figure 6.1 illustrates the process of capturing the live news feed and segmenting the broadcast into semantically coherent news stories.

Between November 2008 and April 2009, we captured the DVB-S signal of these news bulletins and de-multiplexed the stream of the analogue teletext signal and the audiovisual signal. Within this process, the audiovideo signal was converted to MPEG-1 format. Following O’Connor et al. [2001], we used a colour-histogram-based approach to detect shot boundaries in these videos. In the video retrieval domain, shots are usually visualised by static, representative key frames of the shots. In order to determine such key frames, we calculated the average colour histogram for each shot and extract the frames with the shot which are closest to the average. This resulted in a set of key frames for each shot which we then combined to a single animated key frame in GIF format⁶⁻¹.

The next challenge was to combine these shots to larger, semantically related, story segments. The news broadcasts were segmented into stories by individually processing the video and teletext streams. The story segmentation of the video stream was achieved by detecting anchor person shots and the story segments from the text stream were obtained by a semantic Latent Dirichlet Allocation (LDA) based approach. Both individual streams were then combined to identify the story boundaries following Misra et al. [2010]. The segmentation output quality is average when compared with those

⁶⁻¹Graphics Interchange Format

segmentation approaches that have been evaluated within the TRECVideo story segmentation task.

6.2.2 Categorising News Stories

After automatically segmenting the latest news, our next processing step is to classify each news story based on its content. Most news content providers classify their news in accordance to the IPTC standard, a news categorisation thesaurus developed by the International Press Telecommunications Council⁶⁻². We assume that a categorisation of our data corpus using this standard will lead toward a structured user profiling approach. Since the development of an optimal classification scheme is out of scope of this thesis, we rely on OpenCalais⁶⁻³ to classify each story into one or more news categories as defined by the IPTC. OpenCalais is a free Web Service provided by Thomson Reuters, that attempts to identify the “aboutness” of unstructured text documents. It will be introduced in further detail in Section 6.2.3. During the time of the experiment, the latest version of OpenCalais was able to classify text into the following IPTC NewsCodes taxonomy subjects:

- Business & Finance
- Entertainment & Culture
- Health, Medical & Pharma
- Politics
- Sports
- Technology & Internet
- Other

Note that OpenCalais can categorise text documents into more than one subject. Thus, the news stories can be aligned to several IPTC categories.

6.2.3 Semantic Annotation

In a next step, we aim to identify concepts that appear in the stories. Once these concepts have been positively identified, the Linked Open Data Cloud can be exploited to further annotate the stories with related concepts. We have discussed that DBpedia is

⁶⁻²<http://www.iptc.org/>, last time accessed on: 20 December 2009

⁶⁻³<http://www.opencalais.com/>, last time accessed on: 20 December 2009

the backbone of the Linked Open Data Cloud. We hence aim to connect concepts with this ontology; however, various challenges are encountered during this process.

First of all, how can we determine concepts in the story which are strong representatives of the story content? In the news domain, named entities are considered to be strong indicators of the story content, since they convey the most important information among all terms in the news story [Bell, 1991]. Therefore, we extract persons, places and organisations from each story transcript using Thomson Reuters’ OpenCalais Web Service. Iacobelli et al. [2010] argue that this web service is an “off the shelf [...] system that works very well for detecting instances of common categories” in text documents. It relies on natural language processing and machine learning techniques to identify named entities. In addition, it compares identified entity strings with an up-to-date database of entities and their spelling variations to normalise the entity instances.

The second problem is that named entities can be ambiguous. For example, the entity “Scotland” can stand for the country in the United Kingdom or for various towns in the United States. Furthermore, it could be a person’s surname; however, within DBpedia, various concepts referring to any of these “Scotlands” exist. Note that term disambiguation is out of scope of this thesis. We therefore rely on the OpenCalais Web Service to resolve the identity of an entity instance. The service supports three types of entity disambiguation: Company disambiguation, Geographical disambiguation and Product disambiguation. Entities are disambiguated by comparing identified entity instances with all known entities within their database. Surrounding text is used as cues to compute an evidence score for each entity⁶⁻⁴. When a text, for example, contains the entity “Scotland” and also other entities such as “Glasgow”, “Edinburgh” or “Gordon Brown”, the probability of this entity referring to the country “Scotland” increases.

The third problem is how these named entities can positively be matched with a conceptual representation in the Linked Open Data Cloud. Since March 2009, OpenCalais is officially part of the Linked Open Data Cloud. It maps disambiguated entities with a uniform resource identifier (URI), which are further mapped with their representation in DBpedia.

Once the link between the story and the DBpedia graph has been established, DBpedia can be exploited to put each identified entity into its context. As explained in Section 5.3.2, entities in DBpedia are solely nodes in a graph and a semantic hierarchy between most neighboured nodes is defined by the SKOS data model. In order to identify the context of each node, we first extract all neighboured nodes in the graph which represent the category where this node belongs to. The corresponding links

⁶⁻⁴<http://www.opencalais.com/documentation/calais-web-service-api/api-metadata/entity-disambiguation>, last time accessed on: 20 December 2009

are defined by the property “skos:subject”. Furthermore, for each identified category node, we extract all categories that have a semantically broader meaning. These are defined by the property “skos:broader”. In the example given in Figure 5.2 on page 122, the concept “Scotland” is directly linked to the broader categories “British Isles” and “United Kingdom”. We refer to these categories as *Layer 1* categories, since they have a distance $d = 1$ to the initial concept. Further, these broader categories are linked to even broader categories (“Northern Europe”, “Archipelagoes in the Atlantic Ocean”, “European Countries”, “European Union Member States”) with a distance $d = 2$ to the concept “Scotland”. They are hence *Layer 2* categories. Further categories are associated accordingly.

In order to set the entities of the video stories into a broad context, we extract up to four layers of broader categories. Note that not all named entities in the data collection have a concept representation in DBpedia. Furthermore, not all identified concepts are linked to broader categories. An overview of the number of entities, concepts and categories (layers $L_1 - L_4$) in the data collection is given in Table 6.1. As can be seen, there are more categories in higher levels than in lower levels, indicating that on average, each concept is linked with several broader concepts.

TABLE 6.1: Number of entities, concepts and categories in the data collection

# Entities	# Concepts	# L_1 Cat.	# L_2 Cat.	# L_3 Cat.	# L_4 Cat.
10666	8124	42661	76250	115200	145491

Finally, all stories are indexed using MG4J [Boldi and Vigna, 2006], an open source search engine.

6.2.4 Discussion

In this section, we introduced the technical implementation which fulfills the requirements of a personal news video corpus which have been discussed in Section 5.3. We illustrated how we record daily news broadcasts from two independent national broadcasting stations and explain our approach of segmenting these bulletins. Further, we explained how we categorise the segmented news stories into broad news categories based on a well-known news categorisation thesaurus. Moreover, we illustrated how we identify concepts in the news transcripts and explain our method to map these concepts with the Linked Open Data Cloud. For both categorisation and concept mapping task, we rely on the Web Service OpenCalais which is provided by Thompson Reuters. According to Butuc [2009], OpenCalais is one of the Top Ten Semantic Web products of

2009. Identifying concepts within the story transcript allows us to set these documents into a semantic concept. As we have shown in Table 6.1, we extracted the corresponding parent nodes of these concepts over four levels. In the next section, we discuss how these link between the concepts can be used to recommend relevant news stories.

6.3 Recommendation Techniques

In the previous chapter, we introduced a visionary scenario where the fictitious John Citizen interacts with a system that gathers his implicit relevance feedback and generates a “personalised interactive multimedia programme”, consisting of news reports that match John’s interests. Outlining the conditions for such scenario, we argued in Section 5.4 for the generation of a long-term profile where John’s interests, represented by the news stories he interacted with, are stored based on broader news categories. Each of these category profiles contains news stories that John has shown interest in before. We further argued in Section 5.3.2 to exploit external knowledge to set the news video collection in a semantic context. In the previous section, we introduced our approach of employing the Linked Open Data Cloud to identify semantic links between concepts of the news stories.

An interesting research challenge is to exploit these semantic links and hence to recommend more news stories that are semantically related to those news stories which can be found in the user’s profile. As we discussed in Section 2.2.3, two main recommendation approaches exist in literature: collaborative filtering and content-based recommendation. Collaborative filtering relies on information gathered from other users who share similar interests. In Chapter 4, we have introduced a collaboration-based recommendation approach which suggests items that other users, who searched for a similar content, interacted with. Within the scenario which we introduced in Section 5.2, collaboration-based recommendations will not be useful, since only those news stories that others interacted with before will be recommended. Latest news, which were just added to the corpus, will thus not be recommended. Research on content-based recommendation can be traced back to information retrieval [Salton, 1988] and information filtering [Billsus et al., 2002]. It is, according to Adomavicius and Tuzhilin [2005], the most common approach for recommending textual items such as text documents or news articles. In this section, we therefore focus on content-based recommendation techniques.

According to Adomavicius and Tuzhilin [2005], most content-based recommendation approaches identify keywords that can then be used to retrieve related content. Content-based recommendations are therefore items which have been retrieved using a

personalised search query. Applying this approach, we propose to create a search query based on the content of each sub category and to retrieve stories using this query. An interesting research challenge is what should be the content of these search queries. Addressing Hypothesis H_7 that ontologies can be exploited to recommend relevant news documents, we propose to consider the semantic link between news documents when selecting query terms. In Section 6.3.1, we introduce our approach of incorporating the semantic link to generate personalised search queries that retrieve news stories that match the user’s interests. In Section 6.3.2, we introduce other text-based recommendation approaches that can be used to benchmark the introduced semantic recommendation technique. Section 6.3.3 concludes the section.

6.3.1 Exploiting the Semantic Link

In Section 5.3.2, we discussed that John, who shows interest in Mediterranean Sites, might follow a news report about the famous sunset on the Greek Island of Santorini. Any interaction John performs on the news report is stored in his personal user profile. We have proposed to split this user profile based on broad news categories and, further, to cluster its content, resulting in specific sub categories SC containing news stories that John has shown interest in. Assuming that each of the sub categories contains stories that cover one or more (similar) aspects of John’s interest, the content of each sub category can be exploited to recommend more news stories that are semantically related to that sub category. Given the examples in the same section, we argued that linking the transcripts of news stories using a generic ontology can set the news stories into their semantic context. The DBpedia ontology, for example, revealed that the two islands Majorca and Santorini are in the Mediterranean Sea. Assuming that the news story about Santorini is stored in a sub category cluster SC , the semantic link between this story and other stories in the collection could be exploited to recommend John a news story about the beginning of the tourist season on Majorca. A similar idea has already been studied by Richardson et al. [1994], who employ WordNet to identify concepts within text documents and compute semantic distance measures between these concepts and given search queries. Although their results, in terms of precision, are disappointing, the authors argue for the use of such semantic links, attributing the weak performance of their experiment to various limitations of their study. Following their thoughts, we therefore hypothesise (H_7) that the semantic information, provided by an ontology, can be used to recommend related news stories. In the remainder of this section, we discuss our approach of employing DBpedia to generate personalised recommendations.

We have shown in Section 5.3.2 that due to the SKOS attribute, DBpedia is a directed, labelled graph D , consisting of parent and children nodes n that are labelled using globally unique identifiers (URIs). Within DBpedia, each child node can be linked to multiple parent nodes, while each parent node can be linked with multiple children nodes. Thus, $D = (V, A)$ where V is a set of nodes and A is a set of directed edges, defined by the SKOS attribute. In general, children nodes are very specific *concepts* while the parents are broader *categories*. Thus, traversing upwards through the knowledge graph, the more general the parent nodes are. We hypothesise that parent nodes can be exploited to identify other child nodes that are semantically related to the concepts of a news story. We therefore propose to form personalised search queries q_n , consisting of these concepts and categories.

An interesting research question is which concepts and categories should be selected to form such queries. An obvious choice, inspired by pseudo relevance feedback techniques, is to determine the most representative concepts. This could be, for example, the most frequent concept nodes within the sub category cluster. A problem is, however, that some documents within the cluster might contain only a small number of concept nodes. In order to overcome this sparsity, we suggest to apply language modelling techniques for smoothing. We hence identify for every document

$$p(n|SC) = \alpha p(n|SC) + (1 - \alpha)p(n|C)$$

the probability of a concept or category n belonging to the identified subcluster SC . Thus, we also consider the appearance of each node within the whole corpus C to identify the most representative nodes. α is used as prior to smooth the impact of $p(n|C)$. Considering the nodes with the highest probability as representative nodes, we suggest to form a search query consisting of these nodes, combined using “or”. Hence

$$q_n = n_1 n_2 \dots n_l, \text{ where } p(n_1|SC) \geq p(n_2|SC) \geq \dots \geq p(n_l|SC)$$

An important question is how many categories should be considered when formulating a search query. Aiming to study research question Q_5 (“How many concepts should be considered to identify similar news stories to recommend to the user?”), we define the parameter l as the length of the query used to retrieve more content that is similar to the user’s interest.

Another question is how the nodes, coming from different depth within the knowledge graph, should be weighted when generating the personalised search query and how the retrieval results should be ranked. Both query weighting and result ranking are separate

research challenges where the personalised search query is used as an input parameter. Hence, these approaches do not directly rely on preceding methods such as the identification of relevant documents used for query expansion or the generation of personalised search queries. We therefore argue that each problem can be treated separately, i.e. independent from previous measures. [Dudev et al. \[2008\]](#) suggest to determine interest scores that express a user’s interests in certain concepts. They argue that these scores can then be used to infer interest in related concepts. As we have shown in [Table 6.1](#), the number of category nodes increases in each layer of our test collection, indicating that there are more broader categories than specific categories in the collection. Considering the structure of the knowledge graph, we therefore suggest to give a lower weighting to nodes from broader layers, since these nodes are rather general and consequently, their importance fades. Furthermore, we argue that each node having the same depth should have the same weight since this allows a diverse representation of each aspect of the corresponding concept node. Thus, we define each news document in the user profile as a document composed of concept nodes and category nodes of different degree of importance that are arranged in layers surrounding the concept nodes. A similar definition is given by [Zaragoza et al. \[2004\]](#), who refer to title and body of a text document as *document fields*. Computing field-dependent term frequencies, they propose the BM25F ranking function, which allows a different weighting for terms of a document by considering the corresponding field. We discussed the ranking function in [Section 2.1.4](#). Applying their approach, we propose to formulate personalised search queries q_n consisting of l nodes n from each field, combined by “or”, which have been weighted in accordance to the importance of their field and to rank the results using BM25F.

6.3.2 Text-Based Recommendation

After introducing our approach of exploiting the DBpedia ontology to generate personalised search queries q_n , we discuss in this section other, purely text-based approaches that can be employed to recommend related news stories. Ignoring any semantic structure, we identified two different sources which can be used to create a search query:

- *Named Entities*: Named Entities such as “Barack Obama” or “The White House” can be used to avoid noise while automatically expanding search queries, since some of them are very specific. According to [Bell \[1991\]](#), named entities are very important in news documents. Analogue to our DBpedia recommendation, we determine

$$p(e|SC) = \alpha p(e|SC) + (1 - \alpha)p(e|C)$$

the probability of a named entity e belonging to the identified sub category cluster SC and suggest to form a search query q_e consisting of l highest ranked named entities combined using the same concatenation as in (6.3.1).

- *Nouns and Foreign Names*: Some words bear more information than other words, hence providing a higher “content load” than other terms. According to [Lioma and Ounis \[2006\]](#), nouns and foreign names provide most information about the content of a document. Thus, we determine

$$p(nf|SC) = \alpha p(nf|SC) + (1 - \alpha)p(n|C)$$

the probability of a noun or foreign name nf belonging to the identified sub category cluster SC and suggest to form a search q_{nf} consisting of l highest ranked nouns and foreign names using the same concatenation as in (6.3.1).

Aiming to ease comparison of our approaches, we rank results using the classical ranking function Okapi BM25.

6.3.3 Discussion

In this section, we introduced a content-based recommendation approach. We propose to exploit the DBpedia ontology to recommend news stories that match the user’s interests based on the user profile. Further, we proposed two other recommendation approaches that rely on the named entities and nouns and foreign names, respectively. In the next section, we discuss problems that arise when these approaches shall be evaluated using standard evaluation measures.

6.4 Generating Relevance Assessment Lists

In the user profiling scenario which has been outlined in the previous chapter, the user relies on a multimedia recommender system to satisfy his personal information needs. In the previous section, we introduced different recommendation approaches, which can be employed under this scenario. In order to evaluate the performance of these recommendations, user studies are required where users interact with the system over multiple iterations. The constrictions of laboratory-based interactive experiments with pre-defined search tasks do not allow the above scenario, since users will not be permitted to search for the content they are really interested in. Moreover, test collections such as TREC News collections or TRECVideo News videos are outdated, which is a

big drawback for potential user-based evaluation of profiling approaches. Users will behave differently when searching for old news instead of the latest news, hence biasing the outcome of such studies. Sanderson [2006] proposes to create individual, context-specific collections. Using up-to-date test collections can motivate the user to retrieve information they are personally interested in, so they can act more naturally while accessing the data collection.

Various challenges, however, arise when user experiments are based on non-standard test collections. Considering that every participant will be allowed to search for topics of personal interest, no common assessment lists can be created. Participants are unlikely to show interest in the same documents. Further, relevance is relative, which makes pooling of the assessed documents impossible. Even if users are interested in the same topic, they will probably be interested in different aspects and will thus judge the relevance of documents differently [Cuadra, 1967]. Individual assessment lists are therefore needed. One possible way of achieving this is to ask the users to judge relevance for every document in the collection. Considering the size of modern data collections, this approach is not feasible. In order to reduce the manual assessment task, we propose to reduce the number of documents that users have to assess by providing them with subsets of the news collection which match their reported interest.

Aiming to generate such lists, we recruited volunteer assessors. We introduce the suspect group in Section 6.4.1. We first collect personal interests of the group on broadcasted news over several weeks. The interest gathering process is introduced in Section 6.4.2. Based on the participants' feedback, we provide them with a number of news video stories related to their needs and ask them to assess their relevance to their defined interests. Like this, we avoid the situation where users have to evaluate all broadcasted material. The assessment results in individual relevance lists containing users' interests in news topics covering several weeks. Section 6.4.3 discusses this procedure. Section 6.4.4 discusses the assessment task. The experimental documents can be found in Appendix B.

6.4.1 Assessment Group

In order to generate necessary relevance assessment data, we recruited 18 volunteers, further referred to as U1 – U18, with diverse backgrounds using various mailing lists and social networking sites. Since the assessment task is a very tedious work, we allowed each participant to follow their own time schedule. The assessment task was split into two main parts, each part ended with an additional questionnaire where the participants were asked to express their opinion about each part.

6.4. Generating Relevance Assessment Lists

Before the actual assessment, the assessors were asked to fill in an entry questionnaire to provide demographic information. The group consisted of 12 male and 6 females with an average age of 26.4 years. A majority of them holds either an undergraduate or postgraduate degree with a background in IT technologies. Our assessment group corresponds to the most active group in online services [Choicestream, Inc., 2008]. We were first interested to find out which sources they usually rely on to gather latest news. Table 6.2 illustrates their answers. The most named answers they selected were news media websites, followed by television news and word-of-mouth. These replies indicate that the participants accept online news, but also rely on television broadcast.

TABLE 6.2: The assessors' most popular media types used to consume latest news

Media Type	#
News Media Webpages (e.g. BBC iPlayer)	18
Television	10
Word-of-mouth	7
Radio (e.g. in the car)	6
Newsfeeds/Newsletters (e.g. Google Alerts)	4

They are hence the ideal audience for news video recommender systems. Moreover, we were interested whether they follow diverse news topics, a premise for the assessment task. Therefore, they were asked to indicate their interests from a list of broad news categories. Further, they were asked to provide different examples for each category to check how diverse their interest really is.

TABLE 6.3: The assessors' news topics of interest

News Topic	#
Entertainment & Culture	14
Technology & Internet	12
Sports	11
Politics	11
Business & Finance	10
Health, Medical & Pharma	5

Table 6.3 shows their answers. The participants provided an average of 3.9 examples per topic. The results indicate that they show interest in a diverse number of news topics. We hence conclude that they are an appropriate group to base our study on.

6.4.2 Gathering of User Interests

In the first part of the assessment task, we aimed to identify the participants' specific interests for news events. Three assumptions underlie this experimental subtask.

1. We assume that each day, national news media report about the most important news events. More specific, we assume that the BBC reports about this event on their news website⁶⁻⁵. This website is one of the most popular news websites in the UK and well-known for its detailed content. In fact, the website won an online journalism award for their coverage of an event happening within the time period of our data collection⁶⁻⁶.
2. Further, we assume that events with the highest media attention are the most important news events. Apart from “silly season” topics, news media cover stories of general interest.
3. Besides, we assume that “typical” news consumers are mainly interested in the most important news.

In order to identify those stories on the BBC News website which received the highest media attention on that day, we rely on Google News which clusters similar news stories from multiple sources and ranks them based on their popularity [Das et al., 2007]. For each day of our experiment, we retrieved the URL, the headline and a short snippet from the BBC News website as provided by the Google News API. For the assessment task, we generated lists of all retrieved stories, separated by the date and split into blocks of two weeks each. Each list hence contained a maximum of 140 stories (10 stories per day and 14 days). Our participants were asked to mark all stories in each list, seven in total, that they were interested in. For further information, they were also allowed to check the actual website on the BBC server. In a second step, they had to categorise the selected articles into related groups and provide each group with a common label. They were asked to choose rather broad labels for each category without using too general descriptions. This advice aimed at avoiding categories of very specific events which might have appeared only once within the whole time period. Table 6.4 provides an overview of assessed news stories and identified news categories.

The consecutive questionnaire aimed at evaluating their assessment experience. Using Five-Point Likert scales, we first asked them to judge the difficulty of the assessment task. The majority claimed that they found the task very simple. The main difficulty

⁶⁻⁵<http://news.bbc.co.uk/>, last time accessed on: 10 December 2009

⁶⁻⁶<http://news.bbc.co.uk/1/hi/entertainment/8289207.stm>, last time accessed on: 10 December 2009

6.4. Generating Relevance Assessment Lists

TABLE 6.4: Summary of the BBC Online News Assessment Task

	U1	U2	U3	U4	U5	U6	U7	U8	U9
# stories	188	340	117	33	90	178	183	84	157
# categories	19	21	28	10	21	29	17	13	43
	U10	U11	U12	U13	U14	U15	U16	U17	U18
# stories	83	40	157	191	97	38	166	118	127
# categories	68	22	32	18	29	17	46	27	15

they reported was that some news stories could be classified as belonging to more than one category which our interface did not support. Since the assessment task took place a few months after the time period of the data corpus, we were interested if this time difference caused troubles for the participants. We therefore asked the participants to judge different statements on Five-Point Likert scales. Some of the scales were inverted to reduce bias. The assessors stated that before starting the task, they had a general idea of which news events happened in the given time period. Moreover, they claimed that they already knew which kind of stories they were interested in before looking at the collection. As we expected, they claimed that they discovered various news events which they were not aware of before. We assume that this might be partly due to the time difference, but also due to a less intensive following of the news events. The majority did not agree with the statement “I marked various news events as interesting even though I was not interested in them at the given time period”. We conclude that the time difference did not influence the assessors judgment on what they find interesting. The selected categories should therefore be a realistic representation of the assessors’ interests in news within the time period.

6.4.3 News Video Assessment

Knowing the users’ categories of interest, the second part of the experiment aimed at identifying news reports in the video corpus for each category of interest. In an ideal case, the participants would be asked to assess the full data corpus in order to identify these video clips which are relevant to their identified interests. Due to the size of the data collection, however, this approach is not feasible. Hence, it is necessary to provide the participants with a subset of the corpus which they should assess accordingly.

In order to identify a good subset for each category of interest, we exploit a simple observation: Studies (e.g. [Kumaran and Allan, 2004; Bell, 1991]) have shown that named entities such as persons, locations or organisations play a key role in news re-

6.4. Generating Relevance Assessment Lists

ports. The news documents which have been marked and classified in the preceding subtask contain many named entities. Assuming that the same news events which are broadcast have also been reported online, these named entities should also be mentioned in the video report about the same event. Considering that both textual and video news are published by the same news content provider (BBC in our case), it is even more likely that the same terms are used analogically. Moreover, since the textual reports usually contain more details than short video clips, there is a high probability that all named entities which are mentioned by the reporter in the video also appear in the text report. The most important entities from the textual documents should hence provide a good presentation of the content of each category. Further, retrieving news stories using these named entities as a search query should provide a significantly smaller subset of the data corpus which can then be assessed by the participant. Therefore, we use the freely available LingPipe toolkit⁶⁻⁷, at default settings (trained on the MUC-6 English corpus) to extract all named entities from every assessed document. In a next step, we combine the top ten percent most frequent entities of each category of interest using the “or” operator to form a search query.

Using the interface shown in Figure 6.2, the participants were now presented a result list of each category of interest. The label of the category, referred to as an “aspect”, is given on top of the list. Results were ranked using BM25. In addition, each retrieved story had an additional ranking bar where users were asked to assess how much this result is relevant to the given category. Search results were split into several pages containing 15 results each and the participants were asked to assess at least the first three pages. After finishing the assessment for one category, they could click on “Next aspect’s result” on the top of the interface to start the assessment of the next category.

TABLE 6.5: Summary of the News Video Assessment Task

	U1	U2	U3	U4	U5	U6	U7	U8	U9
# days with annotated results	70	76	65	39	50	59	73	78	59
# relevant assessed stories	234	297	217	101	112	155	302	99	203
	U10	U11	U12	U13	U14	U15	U16	U17	U18
# days with annotated results	44	52	69	58	36	51	69	71	32
# relevant assessed stories	156	137	200	187	69	124	187	160	95

Table 6.5 shows the summary of the news video assessment task. As can be seen, the assessment task ended with diverse results, indicated by the different number of relevant assessed stories and different number of days with annotated results.

⁶⁻⁷<http://alias-i.com/lingpipe>, last time accessed on: 10 December 2009


6.4. Generating Relevance Assessment Lists

Logged in as: u1 [Log Off](#) [Next aspect's results \(1/19\)](#)

Results Panel

Results for aspect #1("olympic") Results: 1-15 16-30 31-45 46-60 61-61

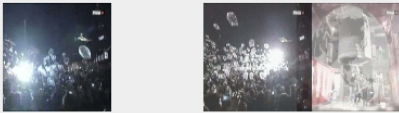
Collapse result "Story #14 from 7th November 2008"



0 1 2 3 4 5

Named Entities: US; Chicago; America
People: Barack; Obama; Adam; Brooks; .Well; Warren; Buffet; Rahm; Emanuel; David; Axelrod
Date: 07-11-08
Broadcaster: BBCOne
Transcript: airways. the us president-elect barack obama will hold his first news conference today after a meeting with his economic advisors. one of the key posts still to be announced is that of treasury secretary, a crucial position given the current financial instability. our correspondent adam brooks is in chicago now. what's likely to come out of this news conference? .well, barack obama is going to be

Collapse result "Story #88 from 31th December 2008"



0 1 2 3 4 5

Named Entities: Tokyo
People:
Date: 31-12-08
Broadcaster: ITV
Transcript: tower. four hours later it was tokyo's turn to bid farewell to 2008. tokyo was the next major city to welcome in the new year.

FIGURE 6.2: News Video Assessment Interface

Figures 6.3 (see page 152) and 6.4 (see page 153) show the numbers of relevant rated stories and the distribution of topics of interest per day for an example user, User 7. Similar patterns can be observed for all participants. As these figures illustrate, the occurrence frequency of topics of user's interest is highly variable. Since users will show diverse interest in news stories on various days, we thus conclude that these assessment lists reflect realistic user interests.

In the final questionnaire, we aimed at evaluating whether the presented subset of the data corpus was appropriate. Using Five-Point Likert scales, we asked the participants to judge whether the displayed news stories were related to the according news aspect. Even though the majority had a neutral perception towards this statement, 43% slightly agreed to it. Moreover, they were asked to judge whether the news stories covered most facets of the corresponding aspect on a Five-Point Likert scale. Again, the participants tended to agree with the statement. We therefore conclude that using the news article assessments to identify good search queries resulted in sensible subsets of the actual video data corpus.

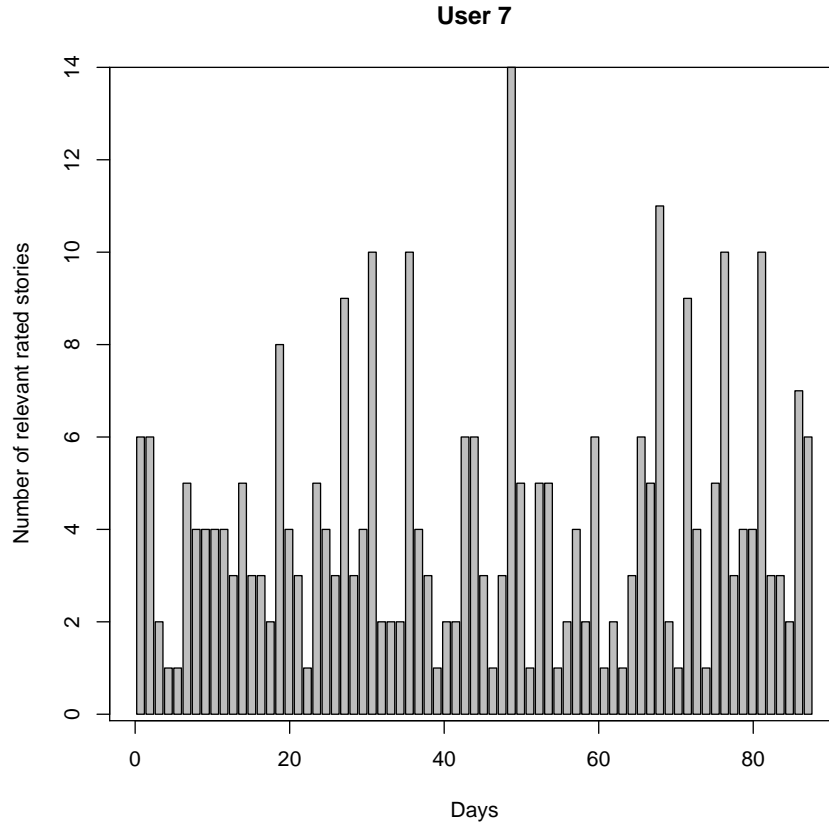


FIGURE 6.3: Number of relevant rated stories per day (User 7)

6.4.4 Discussion

In this section, we introduced an approach of generating personalised relevance assessment lists. In order to reduce the amount of manual labour, we aimed at adapting the assessable documents to the assessors' personal interests. Both quality and quantity of the resulting lists varies from user to user though. While some users provide a large amount of assessments, other users assess a small amount of stories only. Consequently, not all relevant documents are really assessed to be relevant by the users. Nevertheless, since this is a well known problem that also influences other well-established relevance assessment approaches, we conclude that our assessment task resulted in a good representation of users' interests over a longer time period.

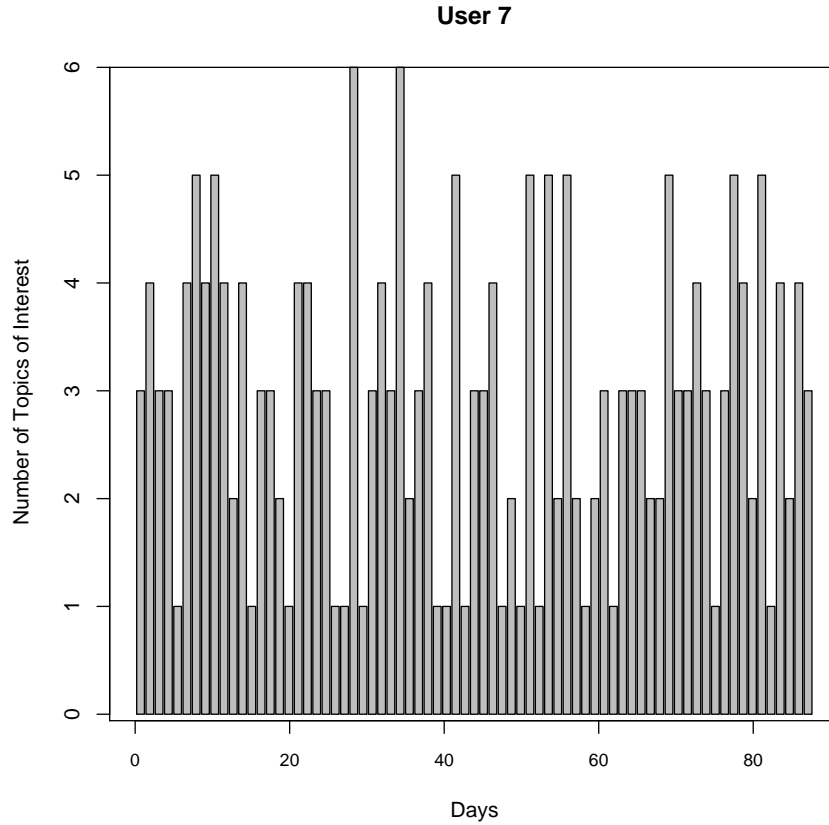


FIGURE 6.4: Number of topics of interest per day (User 7)

6.5 Generating Simulated User Profiles

The relevance assessment lists which have been introduced in the previous section express the interests in news events of 18 potential users of our news video retrieval system. Considering these interests as relevance assessment data fulfils one requirement for a simulation-based evaluation. Another requirement is an implicit user profile of a representative user who interacted with the system over a longer period of time. In this section, we introduce our approach of creating a simulation-based user profile. In Section 6.5.1, we define interaction patterns that may appear when a user interacts with a news video recommender interface. Similar to our approach introduced in Section 3.3, the interactions consist of low-level events. In Section 6.5.2, we determine probability values for the transitions between these states by performing a preliminary user study. Finally, we create implicit user profiles by applying the interaction patterns and identified transition probabilities in Section 6.5.3. Section 6.5.4 discusses our approach.

6.5.1 Training a User Interaction Model

The first step towards evaluating our experimental parameters is to simulate a user interacting with the system. As discussed in Chapter 3, Dix et al. [1993] argue that user interactions in interactive systems can be represented as a series of low-level events, e.g. key presses or mouse clicks. We already introduced possible user actions that are supported by state-of-the-art video retrieval interfaces. As we discussed, some events are independent, while others depend on preceding events. In Chapter 7, we will introduce a news video recommender system that supports the long-term scenario introduced in the previous chapter. Its interface, shown in Figure 7.2 of Section 7.2.2 supports four types of such low-level events:

1. *Previewing*: Hovering the mouse over one of the key frames in the result list pops up a tooltip showing additional information about the news story.
2. *Clicking result*: A click on a result in the result list will expand the according news story and display further information.
3. *Browsing*: A click on any animated shot segment in the expanded view of a news story will centre the according shot. In this way, the user can browse through the shots of a story.
4. *Viewing*: Clicking on the play button in the expanded view will start playing the video.

Two events can be triggered independently from others: Users can always move the mouse over a result to get more information (tooltip event) and can always expand a search result (clicking event). Once a story is expanded, the user can browse through the shots (browsing event) or start playing the video (viewing event). The latter events are hence dependent on the clicking event.

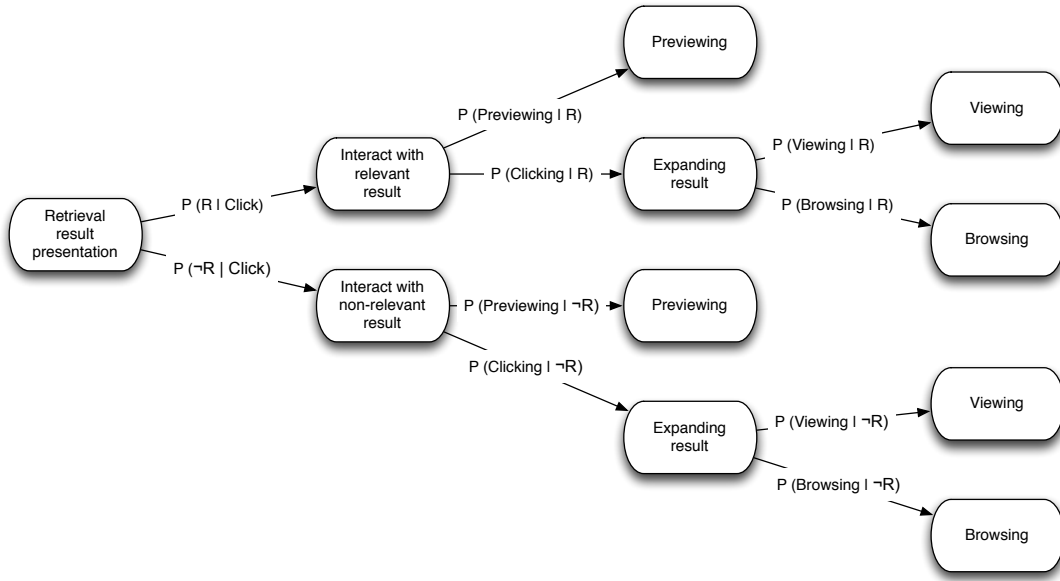
Similar to our approach introduced in Section 3.3, we describe possible event sequences as a Markov Chain. Markov Chains consist of states and transitions. A state change is triggered by a certain event with a certain probability. Figure 6.5 illustrates the possible user interactions of users using the example interface. The probabilities of the above introduced events trigger the transitions between the different states. Note that for simplicity, we consider users doing every event only once.

Transitions are defined as follows:

$$P(R|\text{Click}) = \frac{\# \text{ relevant clicks}}{\# \text{ total clicks}} \quad (6.1)$$

6.5. Generating Simulated User Profiles

FIGURE 6.5: Markov Chain of User Actions



$$P(\neg R | \text{Click}) = \frac{\# \text{ non-relevant clicks}}{\# \text{ total clicks}} = 1 - P(R | \text{Click}) \quad (6.2)$$

$$P(\text{Clicking} | R) = \frac{\# \text{ clicks on relevant stories in result set}}{\# \text{ relevant rated stories}} \quad (6.3)$$

$$P(\text{Clicking} | \neg R) = \frac{\# \text{ clicks on non-relevant stories in result set}}{\# \text{ non-relevant rated stories}} \quad (6.4)$$

$$P(\text{Previewing} | R) = \frac{\# \text{ tooltip highlighting on relevant stories in result set}}{\# \text{ relevant rated stories}} \quad (6.5)$$

$$P(\text{Previewing} | \neg R) = \frac{\# \text{ tooltip highlighting on non-relevant stories in result set}}{\# \text{ non-relevant rated stories}} \quad (6.6)$$

$$P(\text{Viewing} | R) = \frac{\# \text{ playing of relevant stories in result set}}{\# \text{ relevant rated stories}} \quad (6.7)$$

$$P(\text{Viewing} | \neg R) = \frac{\# \text{ playing of non-relevant stories in result set}}{\# \text{ non-relevant rated stories}} \quad (6.8)$$

$$P(\text{Browsing}|R) = \frac{\# \text{ browses of relevant stories in result set}}{\# \text{ relevant rated stories}} \quad (6.9)$$

$$P(\text{Browsing}|\neg R) = \frac{\# \text{ browses of non-relevant stories in result set}}{\# \text{ non-relevant rated stories}} \quad (6.10)$$

Having defined a Markov Chain to simulate user interactions, the next step is now to determine realistic probabilities for each transition in the chain. The best way to simulate realistic user interaction patterns is to analyse how real users interact with the video retrieval system. A statistical log file analysis of this study can then provide an insight into real users' interaction patterns. We therefore performed a preliminary user study to determine representative usage patterns. The study will be discussed in the following section.

6.5.2 Determining Usage Patterns

Since we aim to simulate users interacting with a news recommender system over multiple search sessions, the statistics used to determine probabilities for each transition should be as realistic as possible. We hence had to study users interacting with a system for several days. Besides, we wanted to evaluate the use of the system in a realistic scenario. In the remainder of this section, we introduce a preliminary study where users were asked to include a novel news video recommender system into their daily news consumption routine. Note that we focus in this section on the log file analysis of this study only. A more detailed summary of the experiment is discussed in [Hopfgartner and Jose \[2010a\]](#).

System Description

In this section, we provide a brief description of the news video recommender system used within this study. A thorough description of the system and its data collection is given in Chapter 7. Prior to starting the experiment, we captured the daily news broadcasts and processed the bulletins as described in Section 6.2. Thus, our data collection consists of news stories covering the last six months before the start of the experiment. In addition, we updated the story index every day during the experiment shortly after the latest broadcast to add the latest news stories.

We further implemented an interface which allows the participants of the user study to access this data. The interface will be described in detail in Section 7.2.2. The system monitored the users' interactions with its interface and stored this information in the

user profile as discussed in Chapter 5. The content of the users' profiles is displayed on the navigation panel of the left side of the interface.

Experimental Design

Participants were paid a sum of £10 each to use the system as additional source of information in their daily news consumption routine for up to seven days. Their interactions (e.g. mouse clicks and key presses) with the system were logged to evaluate the approach. They were asked to use the system for up to ten minutes each working day to search for any topic that they were interested in. They were further asked to indicate whenever they found a news story which interested them. In addition, we also created a simulated search task situation as suggested by [Borlund \[2003a\]](#). Our expectation was twofold: First of all, we wanted to guarantee that every user had at least one topic to search for. Moreover, we wanted the participants to actually explore the data corpus. Therefore, we chose a scenario which had been a major news story over the last few months:

“Dazzled by high profit expectations, you invested a large share of your savings in rather dodgy securities, stocks and bonds. Unfortunately, due to the credit crunch, you lost about 20 percent of your investment. Wondering how to react next and what else there is to come, you follow every report about the financial crisis, including reports about the decline of the house market, bailout strategies and worldwide protests.”

Each participant started with an individual introductory session, where they were asked to fill in an entry questionnaire and could familiarise themselves with the interface. Since the system was available online, they could conduct the rest of the experiment from any computer with an Internet connection. This results in an uncontrolled evaluation environment, which we found necessary to evaluate the potential of the system. Every day, they were asked to fill in an online report where they were encouraged to comment on the system as they used it. At the end of the experiment, everyone was asked to fill in an exit questionnaire to provide feedback on their experience during the study.

Participants

16 users (3 female and 13 male) participated in our experiment. They were mostly postgraduate students or postdoctorate researchers with an average age of 30.4 years and a good level of English. They considered themselves as computer experts and indicated

that they often use online search services to retrieve multimedia content. The most popular services they named were Google.com, YouTube.com and Flickr.com. Their favourite sources for gathering information on the latest news stories are news media web portals, word-of-mouth and the television. The typical news consumption habit they described was to check the latest news online in the morning and late at night after dinner. We hence conclude that the participants are familiar with various multimedia search interfaces and rely on both televisual and online media. They therefore represent the main target group for the introduced retrieval system.

By asking for daily reports, our goal was to evaluate the users' opinion about the system at various stages of the experiment. The first question was to find out what the participants actually used the system for. The majority of participants used it to retrieve the latest news, followed by identifying news stories they were not aware of before. Furthermore, we were curious to see what news categories they were interested in. As Table 6.6 shows, the participants followed various news categories. These diverse answers suggest that users did not only use the system to retrieve stories for the pre-defined search task, but also used it for their own information needs, e.g. to follow latest news or to discover other news stories that match their interests.

TABLE 6.6: The participants' news topics of interest

News Topic	#
Politics	91
Business & Finance	74
Technology & Internet	50
Sports	44
Entertainment & Culture	39
Health, Medical & Pharma	27
Other	8

Statistical Log File Analysis

In order to obtain a set of characterisation parameters, we use statistical information of the 16 users to calculate probabilities of users performing certain types of actions. Our first interest is here to judge the quality of the dataset by analysing the number of clicks performed on relevant stories. Since participants of this user study were motivated to retrieve any topic they wanted, story relevance cannot be generalised. What user A might find relevant is completely irrelevant for user B. Therefore, we first determined the probability value $P(R|Click)$ (Equation 6.1) for each individual user, which we then

averaged. According to the log files, the average probability of clicking on a document and rating this document $P(R|\text{Click})$ is 0.55, a rather high value. In other words, approximately every second story that the users interacted with was labelled to be relevant by the according user. Table 6.7 shows the averaged probabilities of an implicit action being performed on relevant and non-relevant using the formulae introduced in the previous section.

TABLE 6.7: Probability values of possible action types

Action Type	Probability
$P(\text{Clicking} R)$	0.34
$P(\text{Clicking} \neg R)$	0.04
$P(\text{Previewing} R)$	0.21
$P(\text{Previewing} \neg R)$	0.02
$P(\text{Viewing} R)$	0.42
$P(\text{Viewing} \neg R)$	0.043
$P(\text{Browsing} R)$	0.97
$P(\text{Browsing} \neg R)$	0.01

6.5.3 Creating User Profiles

Since we want to evaluate the effect of various parameters over a longer period of time for various users, we have to create implicit user profiles for each user. Exploiting the possible user actions and the determined probability values, we create these profiles by simulating the users interacting with the system for every day that has assessed ground truth data. We simulate the following usage scenario:

“Imagine a user who is interested in multiple news topics. They registered with a news recommender system with a unique identifier. For a period of five month, starting in November 2008, they log into the system, which provides them access to the latest news video stories of the day. On the system’s graphical interface, they have a list of the latest stories which have been broadcast on two national television channels. They now interact with the presented results and logs off again. On each subsequent day, they log in again and continue the above process.”

Starting with the first day contained in the individual user’s assessment list, we simulate a user interacting with the news stories of the day according to the introduced user patterns. Each time an event has been triggered, we store this implicit action in the user

profile with the according weighting W as introduced in Section 6.6. In this work, we define a static value for each possible implicit feedback event:

$$W = \begin{cases} 0.1, & \text{when a user browses through the key frames} \\ 0.2, & \text{when a user uses the highlighting feature} \\ 0.3, & \text{when a user expands a result} \\ 0.5, & \text{when a user starts playing a video} \end{cases}$$

The session simulation is repeated iteratively. This results in eighteen individual user profiles containing entries of each day of the data collection with different relevance weighting.

6.5.4 Discussion

In this section, we analysed the user interface of a news recommender system and identified specific feedback events. Moreover, we defined possible user actions, consisting of combinations of these feedback events. Transitions between these events can be expressed in probabilities. Exploiting the log files of a user study, we determined statistical probabilities for each transition and simulated a user using the system over a period of five month. The outcome of this simulation is eighteen user profiles which contain weighted stories of every day in the data collection. In the next section, we discuss how these simulated user long-term profiles can be employed to evaluate the recommendation approaches which have been introduced in Section 6.3.

6.6 Simulation-based Evaluation

In Section 6.3, we argued for content-based recommendations by proposing personalised search queries that have been formed based on the content of each cluster of the user profile. We introduced three different query approaches. The query q_n exploits the semantic relationships between the concepts that have been extracted from the news stories in the profile. We further refer to this recommendation approach as *Semantic recommendation*. Further, we introduced purely text-based recommendation queries that consist on the news stories' named entities (q_e) and nouns and foreign names (q_{nf}), respectively. We further refer to these approaches as *Named Entities recommendations* and *Nouns & Foreign Names recommendations*. The latter two methods are considered to be baseline techniques.

As discussed in the Section 6.1, evaluating recommender systems is a complex pro-

cess and there currently exists no standard evaluation methodology to evaluate recommendation techniques over multiple sessions. The interactive user evaluation paradigm cannot easily be applied since it does not allow users to follow their personal information need over a long time period. We therefore argued for the need of a novel evaluation scheme which is based on user simulation. In the previous section, we introduced an approach of generating artificial user profiles. Exploiting personal relevance assessments of eighteen subjects, we simulated a “typical” user interacting with a news video recommender system over a longer time period. The simulated implicit user profiles hence consist of weighted stories that the simulated users showed interest in at a particular time point.

In this section, we introduce how we rely on these simulated profiles to evaluate the introduced recommendation techniques which have been introduced in Section 6.3. In Section 6.6.1, we first introduce parameters that we want to evaluate. Section 6.6.2 introduces the results of our simulation-based evaluation. Section 6.6.3 summarises and discusses the section.

6.6.1 Evaluation Parameters

Each simulated user profile has been created iteratively. For every day which is covered in the assessment data, new documents have been added, resulting in a daily update of the user profile. In order to evaluate the suggested news recommender approaches (“Semantic”, “Named Entity” and “Nouns/Foreign Names”) with respect to the research questions (Q_4) and (Q_5), we can now compute standard evaluation measures. We focus the evaluation on two parameters: The number s of news stories used for clustering each profile and the number of “terms” l forming each search query q_n , q_e and q_{nf} . Within this study, we study the following values:

- $s = \{4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50\}$
- $l = \{1, 2, \dots, 15\}$

Thus, for each day in each user’s profile, we create the search queries for every cluster in the profile, consisting of s news stories. We then trigger a retrieval using the personalised search query of length l . For each assessed day, we thus have

$$|s| \cdot |l| \cdot |\{\text{Semantic, Named Entities, Nouns/Foreign Names}\}| = 11 \cdot 15 \cdot 3 = 495$$

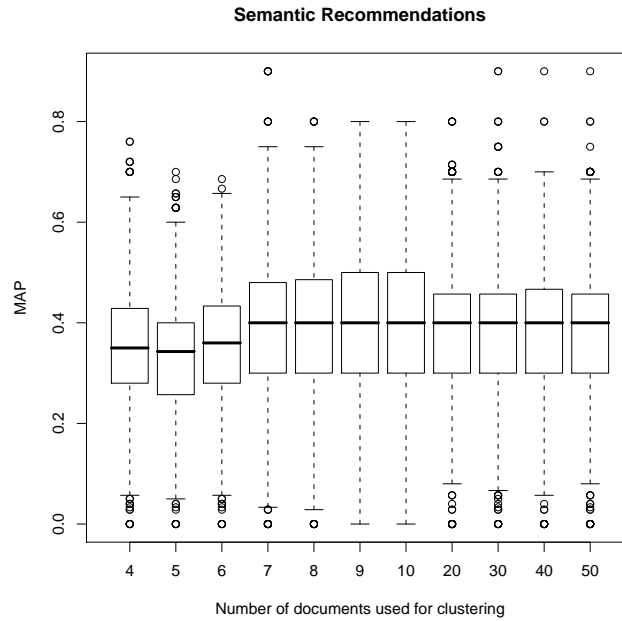
individual runs for every user. By computing standard evaluation measures for every run, we evaluate the quality of the recommendations that the user would get on the

individual day for every aspect of their interest, represented by the clustered content of their profile.

6.6.2 Results

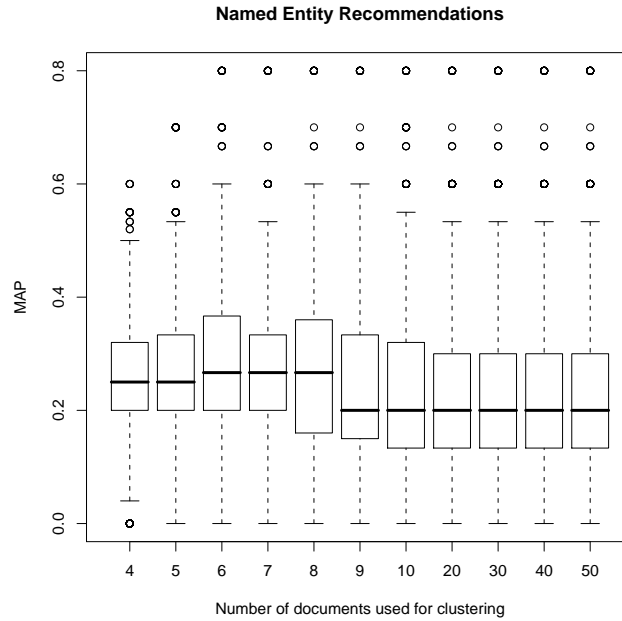
In order to evaluate the first research question Q_4 , we compare the mean average precision of all users for s documents used for clustering. Figures 6.6, 6.7 and 6.8 plot the resulting values for the Semantic run (S), Named Entity run (N) and Nouns/Foreign Names run (NF).

FIGURE 6.6: MAP vs. number of documents used for clustering for Semantic Recommendation run



Various observations can be noted from these figures. First of all, the best performance for all runs can be observed when the search query is based on clusters of 7–10 documents. This suggests that the 7 to 10 highest weighted news stories in a user profile represent best the user’s current interests, answering research question (Q_4). An interesting result is that the parameter s does not influence the performance of the Nouns/Foreign Names run (NF) significantly. This indicates that nouns and foreign names are not optimal to represent the content of a document. The more stories are used to determine the most frequent nouns, the higher is the total number of nouns. The Baseline run exploits this increasing number of nouns and combines the most frequent ones using the “or” operator. The stable performance suggests that the increasing

FIGURE 6.7: MAP vs. number of documents used for clustering for Named Entity Recommendation run

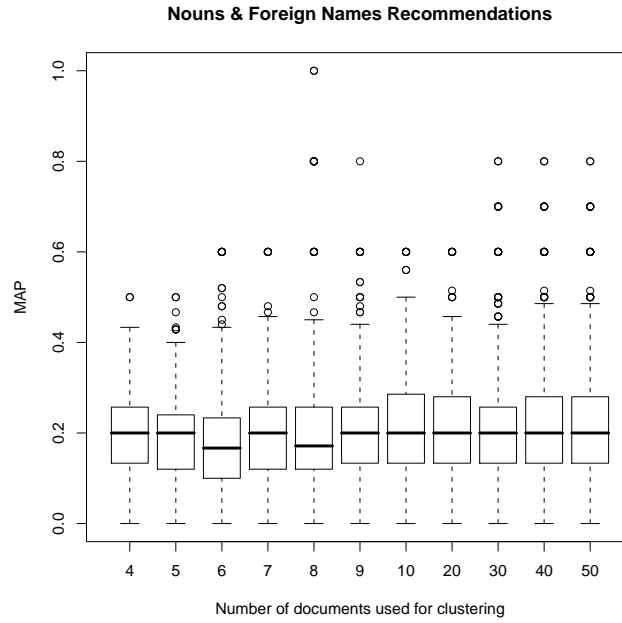


number of nouns does not directly influence the retrieval performance. A reasonable explanation for this is that the most frequent nouns are just not specific enough and hence do not retrieve relevant stories. In comparison, the more specific named entities show a better retrieval performance, suggesting that these, more specific entities, are a better source to create a search query. Both Named Entity (N) and Nouns/Foreign Names (NF) runs are outperformed by the Semantic (S) run, which suggests that exploiting the semantic context of stories in the user profile results in better news recommendations.

All Figures 6.6, 6.7 and 6.8 reveal a large variance for every evaluated parameter. The same observation can be made in Figures 6.9, 6.10 and 6.11 which will be introduced later. We assume that the incoherent quantity and quality of exploited relevance assessment data partly explains this effect. Users show interests in different events to a different extend and at different times. Table 6.5 and Figure 6.4 visualise this diversity. Every user run is based on assessment data of different size and quality and hence influences the outcome of each run.

In order to evaluate the second research question, we compare the MAP of all users for a variable query length l . Figures 6.9, 6.10 and 6.11 plot the according values for the Named Entity run (N), Nouns/Foreign Names run (NF) and Semantic run (S). These figures reveal a minimal or no improvement with longer search queries. A saddle point can be seen at 9–10 query items, suggesting that this might be the optimal query length

FIGURE 6.8: MAP vs. number of documents used for clustering for Nouns & Foreign Names Recommendation run



to identify similar news stories. This would answer research question (Q_5). Again, the Semantic run outperforms both Named Entity run and Nouns/Foreign Names run, suggesting the effectiveness of exploiting the generic DBpedia ontology to recommend related news stories.

An important question is whether this performance difference is significant. Therefore, we performed the Wilcoxon rank-sum test [Wilcoxon, 1945] on the MAP of all runs of every user for every value of query length l and each number s of stories used for clustering. Tables 6.8, 6.9 and 6.10 list the p values of this non-parametric statistical test for a variable number s of stories used for clustering and a constant query length $l = 9$. Note that similar p values can be observed for a variable length of the search query.

Overall, the tables support our conclusions drawn from Figures 6.6–6.11. Using a significance level of 95%, the Nouns/Foreign Names run is, apart from outliers, significantly outperformed by both Named Entity run and Semantic run. Further, in most cases, the Named Entity run is significantly outperformed by the Semantic run. A large performance difference between different users can be noted though. While the semantic-based approaches return significantly better recommendations for some users, it does not provide better recommendations for other users.

Figures 6.12 and 6.13 show a comparison of the recommendation performances,

6.6. Simulation-based Evaluation

FIGURE 6.9: MAP vs. query length for Semantic Recommendation run

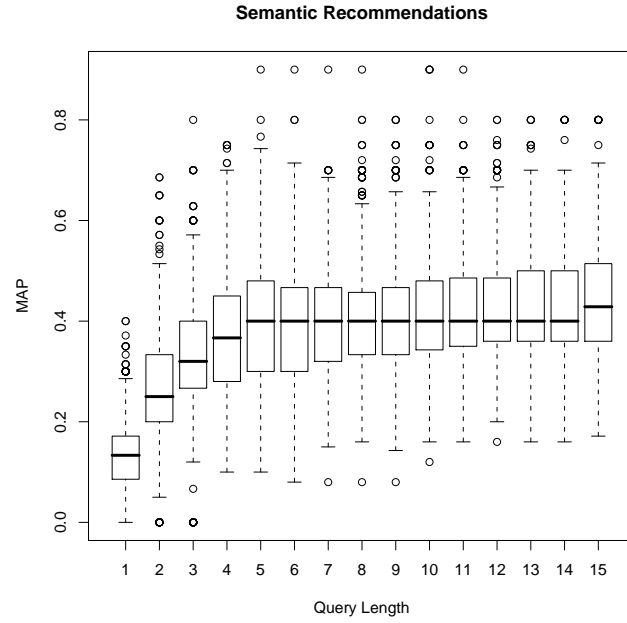


TABLE 6.8: Wilcoxon rank-sum test for variable number of stories used for clustering (Semantic vs. Named Entity Recommendations)

	4	5	6	7	8	9	10	20	30	40	50
U1	0.041	0.056	0.005	0.077	0.014	0.076	0.021	0.024	0.002	0.000	0.000
U2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.009	0.001	0.000	0.000
U3	0.021	0.100	0.259	0.060	0.363	0.808	0.543	0.367	0.015	0.014	0.003
U4	0.025	0.000	0.006	0.002	0.001	0.001	0.000	0.072	0.211	0.515	0.336
U5	0.000	0.000	0.000	0.003	0.009	0.015	0.009	0.000	0.001	0.001	0.001
U6	0.614	0.337	0.337	0.026	0.045	0.060	0.106	0.009	0.201	0.224	0.201
U7	0.065	0.237	0.280	0.060	0.367	0.584	0.377	0.280	0.045	0.242	0.367
U8	0.017	0.073	0.001	0.002	0.001	0.028	0.028	0.138	0.014	0.050	0.014
U9	0.563	0.106	0.138	0.879	0.392	0.279	0.323	0.010	0.288	0.067	0.111
U10	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.016	0.024	0.223	0.616
U11	0	0.004	0.014	0.009	0.011	0.007	0.000	0.000	0.000	0.000	0.000
U12	0.044	0.023	0.005	0.032	0.378	0.108	0.039	0.05	0.299	0.772	0.909
U13	0.002	0.000	0.000	0.007	0.007	0.009	0.003	0.003	0.003	0.015	0.013
U14	0.02	0.007	0.005	0.044	0.098	0.141	0.074	0.041	0.293	0.504	0.589
U15	0.023	0.023	0.018	0.084	0.124	0.15	0.082	0.183	0.292	0.865	1.000
U16	0.000	0.000	0.00	0.001	0.018	0.005	0.032	0.001	0.047	0.056	0.291
U17	0.003	0.062	0.04	0.176	0.424	0.292	0.131	0.238	0.238	0.732	0.780
U18	0.000	0.000	0.000	0.000	0.003	0.002	0.020	0.001	0.005	0.031	0.138

6.6. Simulation-based Evaluation

FIGURE 6.10: MAP vs. query length for Named Entity Recommendation run

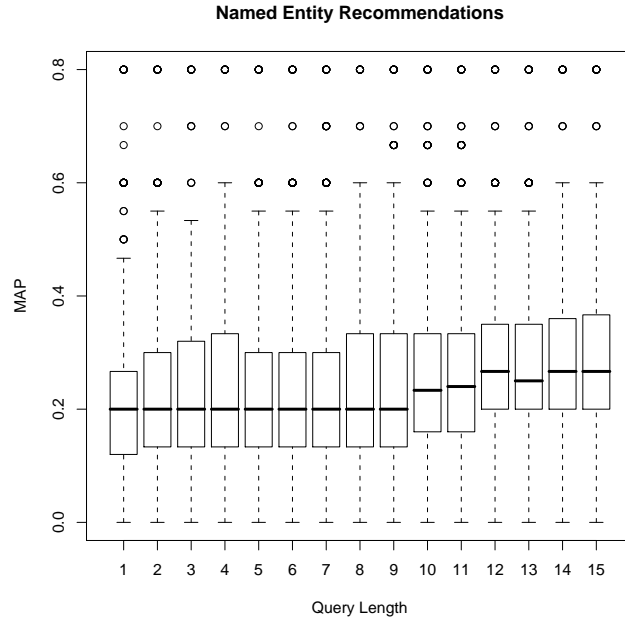


TABLE 6.9: Wilcoxon rank-sum test for variable number of stories used for clustering (Semantic vs. Nouns & Foreign names Recommendations)

	4	5	6	7	8	9	10	20	30	40	50
U1	0.012	0.125	0.045	0.018	0.008	0.007	0.007	0.006	0.121	0.121	0.164
U2	0.002	0.011	0.131	0.084	0.022	0.040	0.006	0.001	0.001	0.001	0.001
U3	0.000	0.000	0.004	0.014	0.025	0.000	0.000	0.000	0.000	0.000	0.000
U4	0.000	0.000	0.001	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000
U5	0.039	0.857	0.465	0.107	0.593	0.142	0.019	0.000	0.000	0.000	0.000
U6	0.001	0.000	0.000	0.007	0.038	0.000	0.004	0.001	0.055	0.011	0.172
U7	0.035	0.137	0.095	0.002	0.018	0.037	0.052	0.002	0.109	0.109	0.151
U8	0.001	0.006	0.245	0.095	0.028	0.212	0.014	0.006	0.003	0.003	0.003
U9	0.190	0.714	0.940	0.009	0.052	0.008	0.003	0.000	0.000	0.000	0.000
U10	0.043	0.022	0.006	0.230	0.296	0.036	0.105	0.002	0.022	0.009	0.119
U11	0.001	0.000	0.003	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000
U12	0.109	0.960	0.704	0.039	0.197	0.034	0.005	0.000	0.000	0.000	0.000
U13	0.003	0.042	0.027	0.048	0.083	0.009	0.005	0.014	0.048	0.030	0.245
U14	0.002	0.001	0.010	0.001	0.004	0.002	0.003	0.000	0.000	0.000	0.000
U15	0.165	0.713	0.846	0.014	0.093	0.004	0.001	0.000	0.000	0.000	0.000
U16	0.770	0.383	0.493	0.429	0.318	0.742	0.419	0.535	0.751	0.821	0.390
U17	0.003	0.038	0.022	0.001	0.002	0.016	0.007	0.001	0.028	0.049	0.049
U18	0.375	0.278	0.326	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000

6.6. Simulation-based Evaluation

FIGURE 6.11: MAP vs. query length for Nouns & Foreign Names Recommendation run

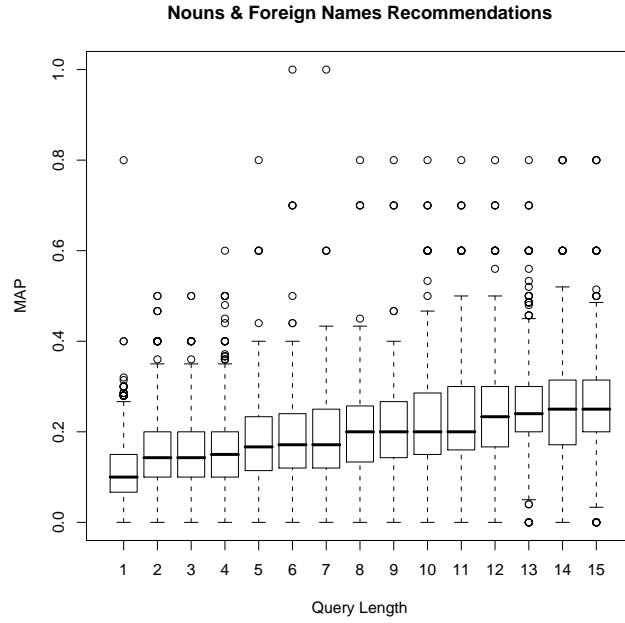


TABLE 6.10: Wilcoxon rank-sum test for variable number of stories used for clustering (Named Entities vs. Nouns & Foreign Names Recommendations)

	4	5	6	7	8	9	10	20	30	40	50
U1	0.065	0.001	0.004	0.025	0.009	0.009	0.034	0.239	0.017	0.003	0.002
U2	0.504	0.253	0.125	0.134	0.323	0.146	0.188	0.032	0.022	0.032	0.032
U3	0.279	0.617	0.025	0.001	0.001	0.144	0.138	0.063	0.053	0.132	0.170
U4	0.001	0.005	0.013	0.002	0.025	0.012	0.310	0.087	0.502	0.431	0.502
U5	0.000	0.000	0.000	0.010	0.038	0.052	0.024	0.048	0.297	0.155	0.195
U6	0.000	0.006	0.018	0.009	0.003	0.030	0.002	0.051	0.009	0.001	0.000
U7	0.007	0.000	0.001	0.003	0.000	0.009	0.004	0.011	0.000	0.000	0.000
U8	0.014	0.004	0.001	0.004	0.008	0.001	0.000	0.000	0.000	0.000	0.000
U9	0.387	0.223	0.175	0.021	0.018	0.743	0.705	0.780	0.689	0.538	0.401
U10	0.004	0.018	0.033	0.011	0.028	0.040	0.452	0.267	0.200	0.101	0.101
U11	0.000	0.001	0.001	0.048	0.206	0.360	0.217	0.063	0.130	0.101	0.130
U12	0.025	0.141	0.294	0.135	0.178	0.073	0.025	0.324	0.771	0.866	0.721
U13	0.904	0.013	0.007	0.014	0.004	0.012	0.029	0.042	0.002	0.005	0.004
U14	0.002	0.001	0.010	0.001	0.001	0.001	0.008	0.086	0.077	0.077	0.077
U15	0.128	0.465	0.083	0.014	0.042	0.956	0.960	0.841	0.861	0.861	0.906
U16	0.352	0.384	0.148	0.263	0.377	0.270	0.379	0.701	0.006	0.002	0.006
U17	0.000	0.001	0.002	0.091	0.103	0.191	0.339	0.579	0.579	0.671	0.671
U18	0.039	0.420	0.239	0.750	0.817	0.402	0.945	0.879	0.036	0.027	0.002

6.6. Simulation-based Evaluation

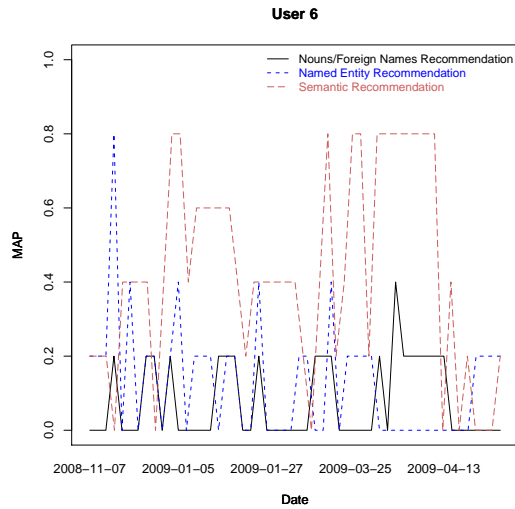


FIGURE 6.12: Recommendation performance of User 6 for every evaluated day with respect to MAP

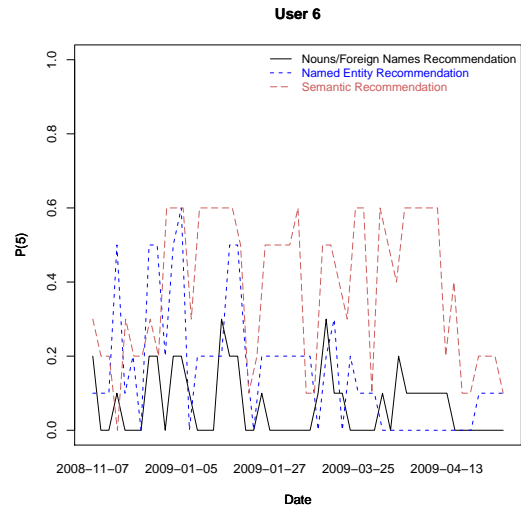


FIGURE 6.13: Recommendation performance of User 6 for every evaluated day with respect to P@5

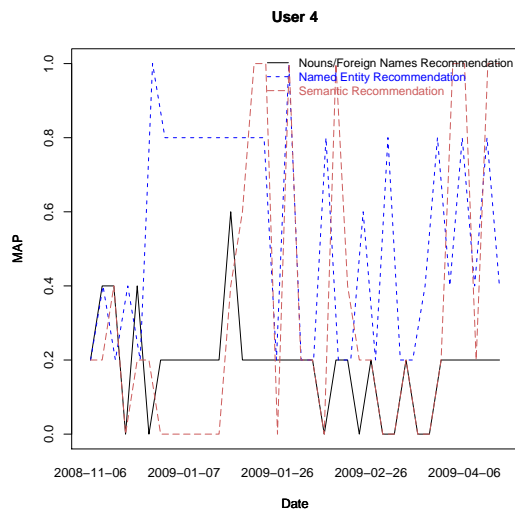


FIGURE 6.14: Recommendation performance of User 4 for every evaluated day with respect to MAP

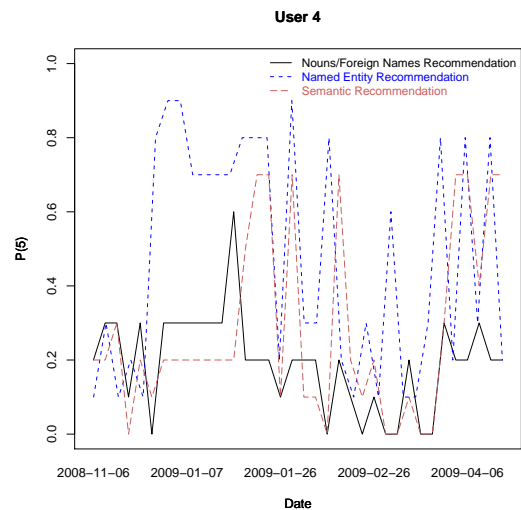


FIGURE 6.15: Recommendation performance of User 4 for every evaluated day with respect to P@5

measured by MAP and P@5, over all days of a representative user (User 6) who significantly benefitted from the semantic-based recommendation. Figures 6.14 and 6.15 show the same comparison for a representative user (User 4) where the Semantic run was not the most successful run. Various conclusions can be drawn from these two figures. First of all, in both cases, the recommendation quality fluctuates massively. The peaks, however, appear synchronously in all runs. As shown in Figure 6.4, a similar fluctuation appears in the assessed list of relevant stories. We therefore conclude that the relevance assessment data directly influences the quality of the recommendation. Moreover, the recommendation quality does not decrease toward the end of each user's profile. Considering that the user profiles are created using implicit relevance feedback, this observation is very interesting. It supports our hypothesis (H_6) that implicit relevance feedback can be successfully exploited to create efficient user profiles.

6.6.3 Discussion

In this section, we employed the simulated implicit user profiles which we introduced in Section 6.5 to evaluate the three recommendation techniques that have been outlined in Section 6.3. For every day that is covered in the simulated profiles, we clustered s highest ranked news stories and created personalised search queries of length l to retrieve other news stories that are related to the simulated user's interest that is represented by the s news stories in the profile. We evaluated these results using standard evaluation measures. The simulations seem to support both hypotheses. The long-term profiles do not illustrate a lower quality of news recommendations after numerous iterations. Hence, we conclude that implicit relevance feedback can effectively be used for automatic user profiling, supporting Hypothesis H_6 . Moreover, the ontology-based recommendations outperform the other comparative runs. Therefore, we conclude that the use of an ontology can lead to better recommendations, supporting Hypothesis H_7 . Further, the simulations revealed that the top 7–10 news stories in the profiles represent best the users' interests, answering research question Q_4 . Answering research question Q_5 , the results suggest that the optimal query length to retrieve relevant recommendations is between 9–10.

6.7 Summary

In this chapter, we evaluated the implicit user profiling approach which we outlined in Chapter 5. One hypothesis we aimed to study was whether implicit relevance feedback can be employed for user profiling. Another hypothesis we aimed to investigate

was whether the selection of concepts in a generic ontology can be used for accurate news video recommendations. Therefore, we introduced our approach of exploiting the Linked Open Data Cloud to set concepts of news stories into their semantic context. We compared this approach with two different baseline runs.

Using a classical evaluation scheme, testing and evaluating these approaches would have been challenging. We therefore argued for the development of a test collection and introduced an approach of generating independent relevance assessment lists that can be used to generate simulated implicit user profiles. In order to reduce the amount of manual labour, we aimed at adapting the documents to assess to the assessors' personal interests. Therefore, volunteers were asked to assess a textual news corpus and to identify news stories they are interested in. Further, they were asked to categorise these news stories into specific news topics. This first assessment step enables us to identify the assessors' interests in news topics. We further exploit this knowledge and identify potential relevant videos in a news video corpus. The assessors were then asked to assess the relevance of this subset. In order to evaluate the recommendation techniques, we proposed a simulation-based evaluation scheme. We defined unique interaction patterns and identified usage patterns by exploiting a user study. Moreover, we employed both patterns and assessment lists to generate implicit user profiles. We then used these user profiles to evaluate our hypotheses and to fine tune various recommendation parameters.

The main conclusion which can therefore be drawn is that the introduced data collection can be used for the benchmarking of long term recommendation approaches. We therefore conclude that our methodology can play an important role in the development of implicit user profiling approaches. Since all results are achieved by employing a simulation, further runs can be performed to fine tune recommendation parameters. Nevertheless, we argue that even though simulations can be used to indicate which retrieval approach is better, it does not replace real user studies. Real users that actually use the system for their own purpose will behave smarter than simulated users. They will, for instance, not just click on random non-relevant news story. Therefore, we conclude that user simulations can be used for benchmarking different approaches, which then have to be confirmed by a successive user study. In the next chapter, we introduce such study.

– *Behaviour is the mirror in
which everyone shows their im-
age.*

Johann Wolfgang von Goethe,
1809

7

Evaluating Semantic User Modelling

In the previous chapter, we introduced a content-based news video recommendation technique that employs a generic ontology to recommend news videos. The recommendation technique relies on an implicit user profiling approach that has been introduced in Chapter 5. We evaluated the ontology-based recommendations by conducting a simulation-based evaluation. In this chapter, we aim to confirm the outcome of this study by performing a user study. In Section 7.2, we introduce the system which we implemented for the study, Section 7.3 discusses the evaluation methodology. In Section 7.4, we show and discuss the results. Section 7.5 summarises the chapter.

7.1 Introduction

In Section 2.3, we argued that evaluation in information retrieval can broadly be categorised into three paradigms. The most dominant one is system-centred evaluation. Indeed, large-scale evaluation campaigns such as TREC are based on it. System-centred experiments are defined by a strict laboratory-based setting. Automatically generated retrieval results are compared to a list of assessed documents and standard evaluation metrics such as precision and recall are computed. The metrics are then used to evaluate the effectiveness of the introduced method [Voorhees, 2005]. Even though system-centred evaluation is suitable for some experiments, it cannot easily be applied to study some research approaches which are focused around the user. This is especially problematic in adaptive information retrieval, which is based on recommending results to

satisfy users' personal interests [Voorhees, 2008].

In the previous chapter, we therefore proposed a simulation-based evaluation scheme that allows us to benchmark an ontology-based news video recommendation technique that exploits implicit user profiles using standard evaluation measures. Even though such evaluation scheme can be employed to fine tune various parameters, we argued that the success of the proposed ontology-based recommendation technique should be confirmed by a user-centred evaluation. As Belkin [2008] pointed out in his keynote speech at ECIR 2008 however, bringing the user into the evaluation process is a grand challenge in evaluating (adaptive) information retrieval (and recommendation) approaches. Kelly et al. [2009] further specifies arising challenges by highlighting that evaluation of information seeking support systems lacks appropriate user and task models as well as test collections. Besides, they stress the lack of longitudinal evaluation designs, a crucial problem when user satisfaction over a longer time period is used as an evaluation measure. As we discussed in Section 2.3, research on adapting content based on users' long term interests has hardly been studied. Few examples include Elliott and Jose [2009], who propose a multi-session study to measure the performance of a personalised retrieval system. In their user study, participants were asked to interact with a personalisation system over multiple days. Users' satisfaction, acquired during different stages of the experiment was used to evaluate their retrieval system. Adopting their approach, in this chapter we evaluate the news video recommendation system which we outlined in the previous chapters by employing a multi-session time series user study. After the quantitative evaluation that has been discussed in the previous chapter, we thus aim to confirm the following hypotheses from a qualitative perspective:

H₆: Implicit relevance feedback techniques can be exploited to create efficient user profiles.

H₇: Ontologies can be exploited to recommend relevant news documents.

The chapter is structured as follows. In Section 7.2, we will introduce the system and recommendation parameters which will be used to evaluate both hypotheses. In Section 7.3, we introduce our methodology of evaluating the news recommendation approach over multiple search sessions. Results of the experiment are discussed in Section 7.4. Section 7.5 summarises the chapter.

7.2 System Description

In order to study the research hypotheses, a user study is required where participants use a news video recommender system over multiple days to satisfy their personal in-

7.2. System Description

formation need. It is well known though that controlled experiments, i.e. experiments in a foreign environment or under someone's supervision can lead to a different behaviour of the test subjects [Campbell and Stanley, 1963]. Aiming to minimize this effect, we wanted to allow the participants of our user study to perform their individual search sessions from a computer of their choice. Therefore, we implemented a Web-based news recommender system based on Asynchronous JavaScript and XML (AJAX) technology. AJAX takes away the burden of installing additional software on each client machine (assuming that a web browser with activated JavaScript is installed)

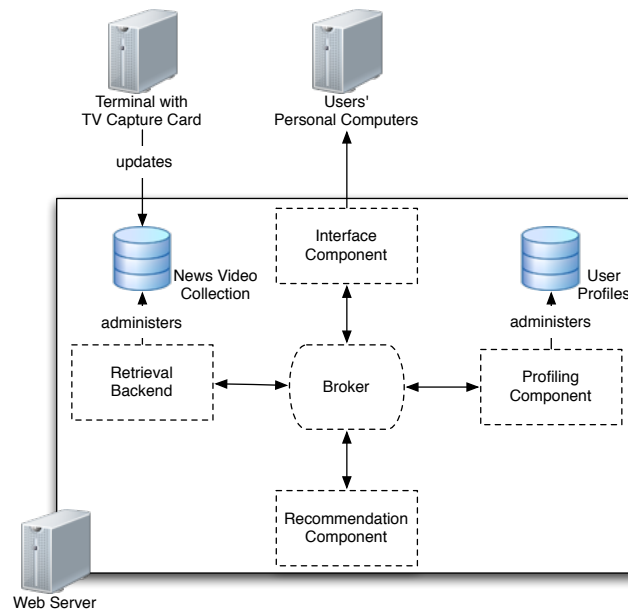


FIGURE 7.1: The conceptual design of the news video recommender system

Figure 7.1 illustrates the conceptual design of the system. As the figure shows, the recommender system can be divided into different components. The first component is the video processing component that is running on a terminal equipped with a TV capture card. This component is triggered every day. Consequently, the component produces an up-to-date video collection consisting of an older news corpus and latest news videos. In Section 7.2.1 we briefly outline the video processing component and the resulting data collection. The main system on the web server consists of four components that communicate with each other through a “broker” component. The data collection is administered by the retrieval backend. We use the already introduced open source full-text search engine MG4J to manage the data corpus. Users access the system using the AJAX-enabled interface component. We introduce this interface in Section 7.2.2. Their interactions are stored in personal user profiles, that are mastered by the profiling com-

ponent. Relevant parameters used within this study are discussed in Section 7.2.3. The recommendation component requires the content of these profiles to generate personalised search queries which are then used to trigger a new retrieval. Aiming to evaluate the ontology-based recommendation technique, we implemented two recommendation techniques: A Baseline Component and a Semantic Component. We discuss both approaches in Section 7.2.4. Section 7.2.5 summarises the system design.

7.2.1 Video Processing Component

In Section 5.3 we discussed the requirements for an implicit user profiling approach with respect to the data collection. The main points we highlighted were that an up-to-date news corpus is required to satisfy the scenario where users browse the corpus to satisfy their personal information need in daily news. Further, we argued in favour of segmenting the captured news bulletins into coherent news stories and suggested to exploit the stories' text transcripts in order to set them into their semantic context.

Differing from the test collection which we introduced in the previous section, the data collection of this user study needed to be updated briefly after the actual broadcast. Thus, we automatically captured both audio-visual and textual streams of the BBC News and ITV Evening News on every week day and segmented, categorised and annotated them as described in Section 6.2. Note though that the version 4.3 of OpenCalais which was used during the experiment supports a richer categorisation of documents as listed in Section 5.3.3. The extended categorisation list⁷⁻¹ includes:

- Business & Finance: topics such as corporate financial results, joint business ventures, global currencies, prices and markets, stocks and bonds, prices, economic forums.
- Disaster & Accident: topics related to man-made and natural events resulting in damage to objects, loss of life or injury.
- Education: topics related to aspects of furthering knowledge of humans.
- Entertainment & Culture: topics such as media, movies and TV, literature and journalism, music, celebrities, entertainment products, internet culture, youth culture.
- Environment: topics related to the condition of our planet such as natural disasters, protection, and their effect on living species as well as inanimate objects or

⁷⁻¹<http://www.opencalais.com/documentation/calais-web-service-api/api-metadata/document-categorization>, last time accessed on: 11 April 2010

property

- Health, Medical & Pharma: topics such as hospitals and healthcare, medical research, diseases, drugs, pharmaceutical industry, health insurance, diet and nutrition.
- Hospitality & Recreation: topics such as eating and travel, leisure/recreational facilities and general activities undertaken for pleasure and relaxation.
- Human Interest: lighter topics of general interest for humans.
- Labour: topics related to the employment of individuals, support of the unemployed.
- Law & Crime: topics relating to the enforcement of rules of behaviour in society, breaches of these rules and the resulting punishments; law firms, legal practice and lawsuits.
- Politics: topics such as government policies and actions, politicians and political parties, elections, war and acts of aggression between countries.
- Religion & Belief: topics such as theology, philosophy, ethics and spirituality.
- Social Issues: topics related to aspects of the behaviour of humans affecting the quality of life.
- Sports: topics such as sports competitions and tournaments, athletes, Olympic games.
- Technology & Internet: topics such as technological innovations, technology-related companies, hardware and software products, internet products and web sites, telecom industry.
- Weather: topics relating to meteorological phenomena.
- War & Conflict: topics related to acts of socially- or politically- motivated protest and/or violence.
- Other: miscellaneous topics not covered by any of the other categories.

After processing the video data, they are uploaded to the web server and the index of the retrieval backend component is updated. Thus, the latest news were available and could be accessed shortly after their broadcasting time.

7.2.2 User Interface Component

In this section, we present the interface that is designed to be used on a PC. It provides various possibilities to supply implicit relevance feedback. Users interacting with it can:

- Expand the retrieval results by clicking on them.
- Play the video of a displayed story by clicking on the “Play video” icon.
- Browse through the key frames.
- Highlight additional information by moving the mouse over the key frames.

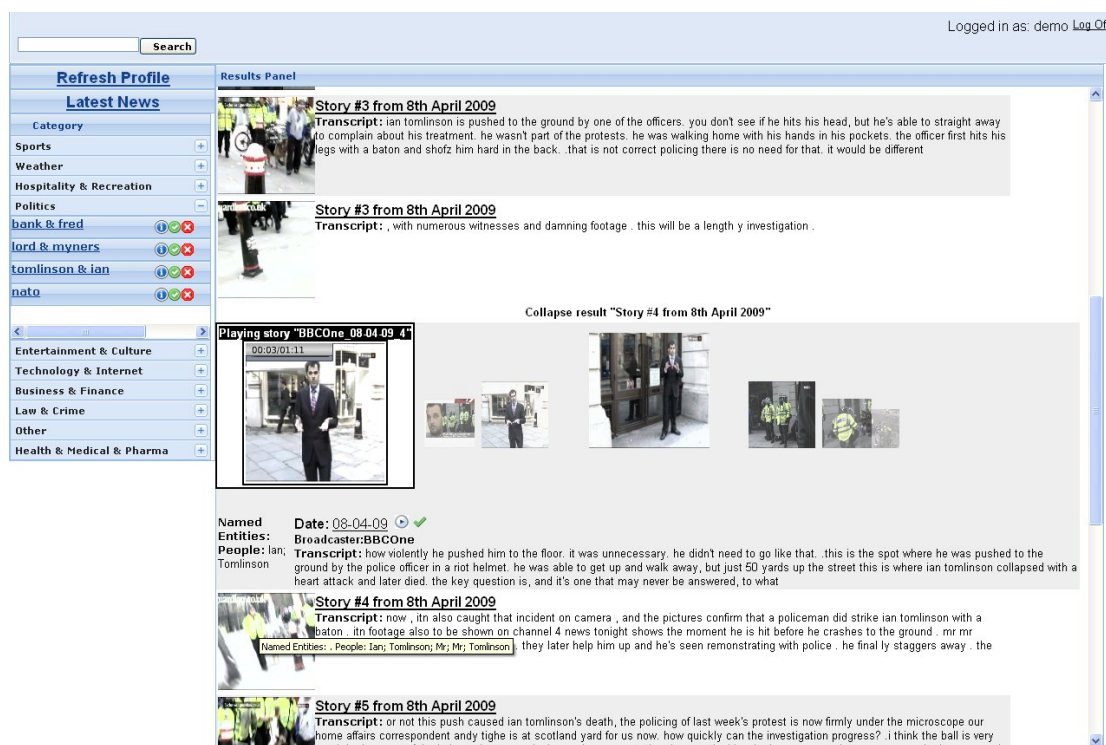


FIGURE 7.2: News Video Recommender Interface

Figure 7.2 shows a screenshot of the news video retrieval interface. It can be split into three main areas: Search queries can be entered in the search panel on top, results are listed on the right side and a navigation panel is placed on the left side of the interface. When logging in, the latest news will be listed in the results panel. Search results are listed based on their relevance to the query. Since we are using a news corpus, however, users can re-arrange the results in chronological order with latest news listed first. Each entry in the result list is visualised by an example key frame and a text snippet of

the story's transcript. Keywords from the search query are highlighted to ease the access to the results. Moving the mouse over one of the key frames shows a tooltip providing additional information about the story. A user can get additional information about the result by clicking on either the text or the key frame. This will expand the result and present additional information including the full text transcript, broadcasting date, time and channel and a list of extracted named entities. In the example screenshot, the third search result has been expanded. The shots forming the news story are represented by animated key frames of each shot. Users can browse through these animations either by clicking on the key frame or by using the mouse wheel. This action will centre the selected key frame and surround it by its neighbouring key frames. Following Furnas [1986], the key frames are displayed in a cover-flow view, meaning that the size of the key frame grows larger the closer it is to the focused key frame. In the expanded display, a user can also select to play a video, which opens the story video in a new panel. The user's interactions with the interface are exploited to identify multiple topics of interests. On the left hand side of the interface, these interests are presented by different categories. Clicking on any of these categories in the navigation panel will reveal up to four sub categories for the corresponding category. Clicking on any of these sub categories results in the generation of a personalised search query exploiting the content of the sub category. The content of each sub category cluster can be edited by clicking on the icon next to the panel's label. The query generation process will be introduced in Section 7.2.4. The profiling component will be introduced in the following section.

7.2.3 Profiling Component

We introduced in Section 5.4 our approach of generating individual implicit user profiles consisting of news stories that users' provided implicit relevance feedback on. The interface which is used within this study provides four features that can be used as implicit indicators of relevance. Similarly to the simulation-based evaluation of Section 6.5.3, we define the following weighting for these features:

$$W = \begin{cases} 0.1, & \text{when a user browses through the key frames} \\ 0.2, & \text{when a user uses the highlighting feature} \\ 0.3, & \text{when a user expands a result} \\ 0.5, & \text{when a user starts playing a video} \end{cases}$$

We further discussed that for the purpose of exploiting these profiles, i.e. to generate recommendations for the users' interests in different aspects of news, the profiles should be

split based on broader categories. We argued that these categories could be broad news topics such as “Sports”, “Politics” or “Business”. While business-related news stories, for example, should be stored in the “Business” profile, other stories should be stored in their respective profiles. Moreover, we argued that a requirement for this approach is to categorise all news stories in the corpus. In Section 5.3, we therefore introduced our categorisation approach using the Web Service OpenCalais. In Section 7.2.1, we have shown that the current version of OpenCalais is able to categorise text documents into eighteen news categories. Within this study, the users’ interests can be split into eighteen separate categories. When a user interacted with news stories of the relevant category, it will be displayed on the left hand side of the interface shown in Figure 7.2.

Since these broad news categories cover several news topics that users can show interest in, we argued that the content of these categories should be clustered into n sub categories that represent users’ interest. In the preliminary user study that we have outlined in Section 6.5.2, the participants named on average 3.8 different sub categories for every broader news category they showed interest in. We therefore set $n = 4$, resulting in four sub categories for every category.

An interesting question is how these sub categories should be labelled to allow the users of our system to get an idea of what to expect when clicking on the appropriate sub category in the interface. We provide sub category labels by extracting the two most frequent named entities of the cluster. Recently, more advanced approaches have been introduced to identify cluster labels though, e.g. by using Wikipedia [Carmel et al., 2009].

As explained before, clicking on one of the labelled sub category panels will trigger a personalised search query that retrieves content-based recommendations for the corresponding sub category. Within this study, we compared two different recommendation approaches that will be introduced in the next section.

7.2.4 Recommendation Component

The most important component within this study is the recommendation component. As is common in information retrieval experiments, the performance of a novel methodology, or approach is usually evaluated by benchmarking various settings and comparing it with a state-of-the-art approach, referred to as the *Baseline*. Following this approach, we define two different recommendation components that will be compared within this study.

Semantic Recommendation Component

The first component, denoted the *Semantic Recommendation Component*, exploits the semantic link provided by the DBpedia ontology, as discussed in Section 6.3.1, to generate a personalised search query q_n . As discussed, q_n is composed of l nodes from different layers within the ontology, combined using boolean “or.” Each layer is treated as an individual field and retrieval results are ranked using BM25F.

Baseline Recommendation Component

In Section 6.3.2, we introduced two text-based search queries, consisting of Named Entities (q_e) and Nouns/Foreign Names (q_{nf}), that can be employed as a potential *Baseline Recommendation Component*. The results of the simulation runs (Section 6.6) suggest that Named Entities result in better recommendations than Nouns and Foreign Names. Consequently, we define the Named Entity run (N) as our Baseline System. Thus, the personalised search query q_e , consisting of l most representative named entities combined by “or” is used to retrieve recommendations, which are then ranked using BM25.

Recommendation Parameters

In Chapter 6, we already benchmarked various parameters to improve the performance of our recommendation techniques. From the simulation, we concluded that the personalised search query q used to retrieve content-based recommendations for the corresponding sub category cluster should be composed of 9–10 query elements. Thus, within this study, we set the query length to $l = 9$. Another result of the simulation was that the 7–10 highest ranked news stories in the user profile represent best their interests. Thus, we set the number of news stories used for clustering to $s = 10$.

7.2.5 Discussion

In this section, we introduced the news video recommender system that is used to evaluate the two research hypotheses that have been outlined in Section 7.1. As we have shown, the system can be segmented into six main components. Every day, one component, the video processing component, captures, segments, categorises and annotates the daily news broadcasts from two major television channels. Users can explore the data collection using a web-based graphical user interface. Since the system is available online, users can access it from any computer with internet access. Their interactions with the system are stored in categorical user profiles that are maintained by the profiling component. We introduced two different recommendation components, a Baseline

Recommendation Component relying on textual features and a Semantic Recommendation Component that exploits the generic ontology DBpedia to recommend further news stories that match the users' interest as represented in their profiles. Parameters of the recommendation technique are set based on the outcome of the simulation-based evaluation in Chapter 6. In the next section, we will discuss the experimental methodology to evaluate the profiling and recommendation approach.

7.3 Experimental Methodology

Aiming to assess the recommendation approach over multiple search sessions, we performed a between subjects multiple time series study similar to the study presented by Elliott and Jose [2009]. Section 7.3.1 provides an overview over the experimental design of the study. In Section 7.3.2, we introduce the data collection which was used for this experiment. In Section 7.3.3, we introduce the participants of our evaluation. Section 7.3.4 discusses the experimental methodology.

7.3.1 Experimental Design

Participants, recruited using department-wide mailing lists, were paid a sum of £15 to take part in our evaluation.

The experiment started with a short introduction in the experimenter's office, where the participants got familiarised with the experiment. After filling an Entry Questionnaire we introduced them to the news video recommender system in a 10-minute training session.

We split the experiment into ten sessions that the participants could perform from a computer with internet access of their choice. They were asked to include the news recommender system in their daily news consumption routine and interact with the system for a minimum of 10 minutes each day. Half of them, randomly assigned, were asked to interact with the Baseline system while the other half had to use the Semantic system. Every day, they received an email that reminded them to continue with the experiment. The participants were told to use the system to explore any news they were interested in. Furthermore, as proposed by Borlund [2003a], we created a simulated search task situation that they could search for in case they did not find any other news of their interest. Our expectation was twofold: First of all, we wanted to guarantee that every user had at least one topic to search for. Moreover, we wanted the participants to actually explore the data corpus. As we will show in Section 7.3.2, various sport events happened during the time of the experiment. We therefore chose a sports dominated

scenario:

“You just started a new job in a remote town away from home. Since you do not know anyone in town yet, you are keen to socialise with your colleagues. They seem to be big sports supporters and are always up for some sports related small talk. Sport hence opens great opportunities to start conversations with them. Luckily, there are various major sports events and tournaments this month which they all show interest in, e.g. the Winter Olympics in Vancouver, the Rugby Six Nations Cup and European Football Tournaments. Every day, you eagerly follow every news report to be up to date and to join their conversations.”

Indicative Request: You should use the recommender system to follow sports related news. This might include major international events such as the Winter Olympics, European football competitions or the Rugby Six Nations cup. Reports might be summaries of the competition day, feature reports about Team GB, or summaries of football/rugby matches. Keep in mind that you should follow the news well enough to be able to chat and socialise with your new colleagues.

We wanted to evaluate the recommendation approaches by comparing the participant’s opinions about the systems during various stages of the experiment. [Campbell and Stanley \[1963\]](#) advise against repetitively prompting the same questions within a short period of time, since the users’ behaviour may adapt based on the intention of the questionnaires. Therefore, we asked them every second day of the experiment, to fill in an online Interim Questionnaire. At the end of their tenth search session, they were asked to fill in an online Exit Questionnaire. Table 7.1 depicts the experimental schedule. All questionnaires and information sheets can be found in Appendix C.

TABLE 7.1: Experiment schedule

Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
System Tutorial	Search Session	Search Session	Search Session	Search Session	Search Session	Search Session	Search Session	Search Session	Search Session
Entry Questionnaire	Interim Questionnaire		Interim Questionnaire		Interim Questionnaire		Interim Questionnaire		Interim Questionnaire
Search Session									Exit Questionnaire

Note that due to the uncontrolled nature of the experimental setting, we had no influence on when and how the participants performed their search sessions. Consequently, the experiment lasted roughly one month, since some participants skipped

various dates, hence increasing the overall duration of the actual experiment. Nevertheless, even though this was not intended, we argue that such user behaviour resulted in more realistic data. Most users check news whenever they feel like and not necessarily every day, as initially intended by us. Further, we argue that skipping days within the experiment can be an interesting challenge for our user profiling approach.

7.3.2 Data Corpus

As discussed in the previous chapters, the news video recommendation approach relies on an up-to-date news video corpus which will be updated constantly. Prior to starting the experiment, we recorded the daily news broadcasts from BBC One (One O’Clock News) and ITV (Evening News) for various months and processed the bulletins as outlined in Section 7.2.1. During the experiment, we automatically updated the corpus by recording and processing the latest news broadcasts from these channels. The participants could hence explore the latest news and access older news. The following events were planned for the time of the experiment:

- The *XXI Olympic Winter Games* in Vancouver, British Columbia, Canada. Athletes from over 80 nations participated in this multi-sport event. The UK was represented by 52 athletes, forming Team GB. With Britain being the host of the next Olympics, we expected a high media attention for the games.
- The first leg of the Round of 16 of the *UEFA Champions League*. The Champions League is an annual football cup competition organised by the European Football Associations (UEFA). It is one of the most watched annual sporting events worldwide. Three UK-based (English) clubs competed in this knockout phase. Considering the high popularity of English Premier League teams all over the UK, we expected various reports about these games.
- The first and second leg of the Round of 32 of the *UEFA Europa League*. The Europa League is another annual football competition organised by UEFA. Six games involved the participation of an English football team.
- *RBS Six Nations Rugby Championship*, an annual international rugby union competition. It is the most popular rugby union tournament in the Northern Hemisphere. Considering the popularity of this sport in the UK, four out of six participating teams came from the British Isles, we expected a high media attention.

Other, non-scheduled breaking news were:

- Armed forces storm Niger presidential palace.
- An 8.8 magnitude earthquake occurs in Chile.

The reported events provide different conditions for our user profiling and recommendation techniques. The Winter Olympics, for example, took place on every day during the experiment. The profile of a user showing interest in the Games would consequently contain many news stories about the Games. The football and rugby games were less frequent. A user interested in these games would hence have less interaction with corresponding reports. Moreover, some events, the “breaking news”, occur all of the sudden and are therefore not represented in the profiles yet. The political situation in Niger, for example was rarely covered by British media before the mentioned incident.

7.3.3 Participants

24 users (16 male and 8 female) participated in the experiment. 21 of them were either Graduate Students or Faculty/Research Staff, three were undergraduate students. The majority of the participants studied or worked on Computer Science related topics, mostly focusing on Human Computer Interaction and Information Retrieval. They claimed to have a high expertise of English with 25% of them being native speakers of English. The majority of 62% of the participants were between 26–30 years old. 24% were between 31–38 years old and three participants were in the age group of 18–25 years.

Multimedia Expertise

As part of the Entry Questionnaire, we were interested in the participants’ expertise when dealing with multimedia content. We therefore first asked them to indicate which online search services they use to retrieve multimedia content. The most common results, which they could select from a list, are shown in ranked order in Table 7.2. Considering the intended use of these online services, i.e. YouTube’s focus on videos and Flickr’s focus on pictures, we note that the participants often retrieve videos and pictures online. Thus, we conclude that they have a high expertise when dealing with multimedia data.

With the next question in the questionnaire, the participants were asked to indicate on a list which media type they usually use to receive latest news. Their replies that have been listed in Table 7.3 indicate that the participants mainly rely on News Media Websites such as the BBC iPlayer portal and the television. We thus conclude that the participants prefer to consume news in video format.

7.3. Experimental Methodology

TABLE 7.2: How the participants retrieve multimedia content

News Source	#
Google (http://www.google.com/)	24
YouTube (http://www.youtube.com/)	23
Flickr (http://www.flickr.com/)	8
Yahoo (http://www.yahoo.com/)	5
Bing (http://www.bing.com/)	4
Other	5

TABLE 7.3: The participants' most popular media types used to consume latest news

Media Type	#
News Media Webpages (e.g. BBC iPlayer)	22
Television	12
Newsfeeds/Newsletters (e.g. Google Alerts)	8
Radio (e.g. in the car)	8
Word-of-mouth	8

News Consumption

In another question, we aimed to identify the participants' main interests. Their results are shown in Table 7.4. They indicate that the participants show interest in a diverse set of news stories, which is a challenge for our user profiling approach.

TABLE 7.4: The participants' news topics of interest

News Topic	#
Technology & Internet	20
Sports	18
Entertainment & Culture	16
Politics	14
Business & Finance	6
Health, Medical & Pharma	7
Other	2

Further, we asked the participants to outline their usual news consumption habits. The majority reported that they follow news during the afternoon or late at night. The majority of the participants reported a daily consumption of news. A common response on how they consume was that they "check it online", e.g. during dinner or their tea break.

Summarising, we conclude that the participants show a high expertise when dealing

with online multimedia content, a pre-condition for our user study. Further, the entry questionnaire revealed that they show diverse interest in multiple topics, which will be a challenge for our user profiling approach. Finally, their responses indicate that they mostly use online services to retrieve latest news. They therefore represent the main target group for our news video recommender system.

7.3.4 Discussion

In this section, we outlined the experimental methodology which we applied to evaluate the long-term user modelling and news video recommendation approach. We first introduced the experimental design of our between subjects multiple time series study. We discussed that users should use our system for up to ten days to follow the latest news. We then introduced the data collection which we used for the evaluation. As we have shown, the data collection consisted of a set of news video reports that had been broadcasted before the start of the experiment. Besides, we automatically updated the corpus every day by adding the latest news broadcasts. We highlighted the major events that happened during the experiment and argued that the news corpus covered multiple news topics and pose different challenges for our experiment. Finally, we introduced the group of users that were paid to participate in our study. Summarising their experience in handling multimedia content and their general interests in news, we argued that they represent the target group of our news video recommender system.

7.4 Results

In order to evaluate the previously introduced hypotheses, we followed a user-centred evaluation scheme where the users' satisfaction and interaction are the most valuable evaluation measures. By asking for frequent reports every other day of the study, our goal was to evaluate the users' opinion about the system at various stages of the experiment. Further, tracking their interactions in log files allows us to get an insight into their activities. In this section, we present an analysis of these questionnaires and the created log files. First, we evaluate the general system usability in Section 7.4.1. Section 7.4.2 summarises the users' judgements and feedback with respect to exploiting implicit relevance feedback for user profiling. Section 7.4.3 focuses on evaluating the quality of the provided recommendations. Section 7.4.4 discusses the results.

7.4.1 System Usage and Usability

The first question of our interim questionnaire was to find out what the participants actually used the system for. We therefore asked them to check on the online form those pre-defined tasks that described best their activity. A summary of their responses over all days is given in Table 7.5. Note that people could select more than one checkbox, hence why percentages add up to more than 100%.

TABLE 7.5: What the system was used for

	Total	Percentage
Finding videos of older news	27	23%
Identifying latest news	110	95%
Identifying news stories you haven't heard of before	52	45%
Other	4	3%

As can be seen, the majority of participants used it to retrieve the latest news, followed by identifying news stories they were not aware of before. Furthermore, we were curious to see what news categories they were interested in. The participants were therefore asked to check in the online questionnaire the corresponding news categories they were interested in during the last days.

TABLE 7.6: News categories that the users were interested in during the experiment

	Total	Percentage
Business & Finance	29	25%
Entertainment & Culture	42	36%
Health, Medical & Pharma	26	22%
Politics	62	53%
Sports	78	67%
Technology & Internet	33	28%
Other	8	7%

As Table 7.6 indicates, the participants followed various news categories. These diverse answers suggest that users did not only use the system to retrieve stories of the pre-defined search task, but also used it for their own information needs, i.e. to follow latest news or to discover other news stories that matched their interests.

Aiming to evaluate the users' satisfaction while interacting with the interface, we asked the participants to judge various statements on a Five-Point-Likert scale from 1 (Agree) to 5 (Disagree). The order of the agreements varied over the questionnaire to

7.4. Results

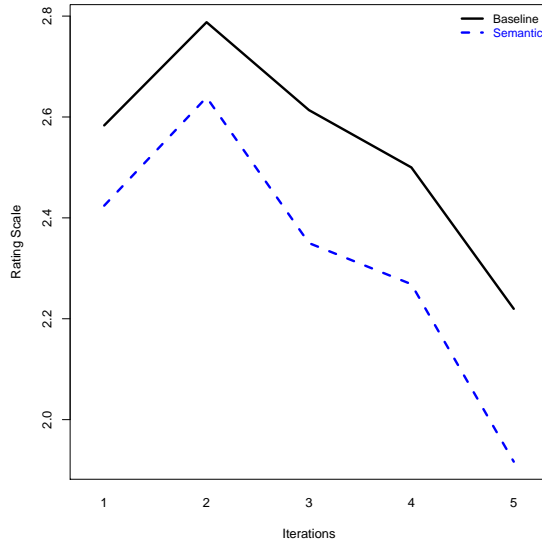


FIGURE 7.3: The interface helped me to explore the news collection (lower is better)

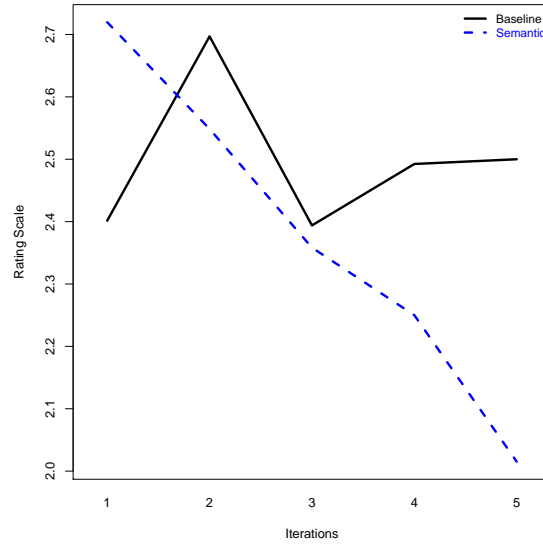


FIGURE 7.4: The interface helped me to explore various topics of interest (lower is better)

reduce bias. Aiming to determine the general usability of the system, we asked them to judge the following statements: (1) “The interface structure helped me to explore the news collection” and (2) “The interface helped me to explore various topics of interest”. Figures 7.3 and 7.4 show the average judgements of all users over all ten days.

Interestingly, the two different user groups had a different perception of the accessibility of the collection and topics of interest. Considering that both groups interacted with the same interface, we assume that the participants generalised their judgments with respect to the whole system they used rather than the interface only. Figure 7.3 shows the users’ agreement that the system helped them to explore the news collection. Neglecting a bump at the second iteration, i.e. the fourth day of the study, a clear trend towards positive perception can be observed. This trend can be explained by the human learning abilities. Once the users got used to the system, they appreciated its functionalities. The bump at the second iteration might be explained by a bug that occurred during the fourth day of the study: even though categories and search results were displayed, the interface did not allow the users to access any sub categories. The same bug can explain the bump that is shown in Figure 7.4, depicting the users’ opinion about the systems’ usability to explore various topics of interest. Remarkable is the better assessment of the semantic-based system which suggests an overall better performance of this recommendation technique. Note, however, that a Wilcoxon rank-sum test did not reveal any significant difference between the user groups.

7.4.2 Exploiting Implicit Relevance Feedback

Figure 7.5 provides an overview over the average number of implicit relevance feedback that the participants provided while interacting with the system. The figures, extracted from the log files of the study, illustrate that users performed a vast amount of interactions that were used to identify their interests. Note that overall, the user group interacting with the semantic recommender system provided more implicit relevance feedback than the users of the baseline system. Moreover, a high activity can be spotted on the first day of the study by the semantic user group. A closer look at the log files reveals that one user browsed through all key frames of the news stories he retrieved. Considering that he performed this action on the first day only, we consider this as an anomaly. Another interesting observation is the decreasing amount of feedback that the users of the baseline system provided at later stages of the experiment. This could indicate that they lost interest in the study, maybe due to inefficiency of provided recommendations.

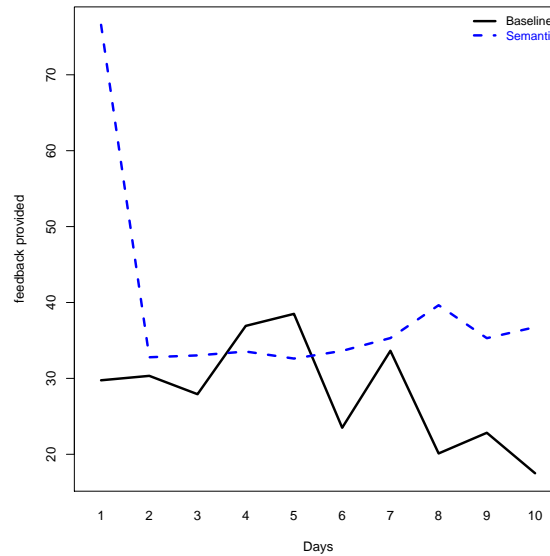


FIGURE 7.5: Implicit relevance provided by both user groups

With the aim of studying our first hypothesis that this implicit relevance feedback can be used to create implicit user profiles, we were interested whether the system was effective in automatically identifying the users' interests. Therefore, we asked the participants to judge the following statement: "Categories were successfully identified and displayed on the left hand side of the interface". Their judgements are depicted in Figure 7.6. Further, we aimed to understand whether the content of these categories matched the users' interests. Figure 7.7 illustrate their judgements of the statement "The displayed sub categories represent my diverse interests in various topics".

7.4. Results

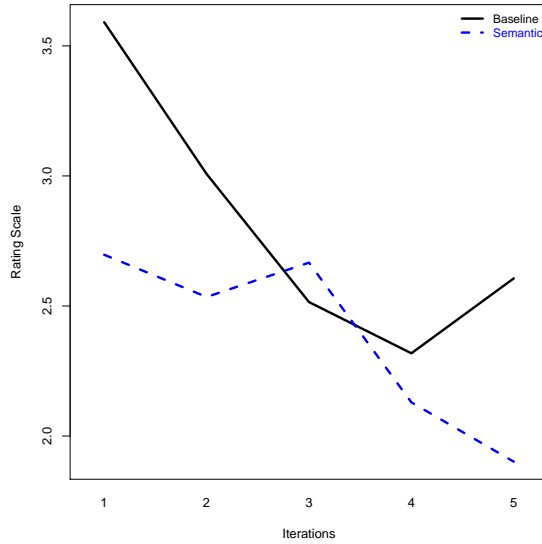


FIGURE 7.6: Categories were successfully identified and displayed on the left hand side of the interface (lower is better)

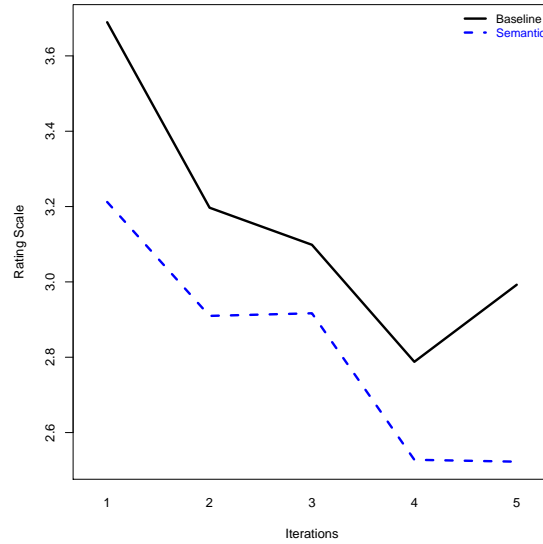


FIGURE 7.7: The displayed sub categories represent my diverse interests in various topics. (lower is better)

As can be seen, the average ratings from both groups indicate a positive tendency towards the two statements. Further, it can be seen that the initial assessment in the first iteration is rather negative, i.e. above the mean of 3. Considering that profiling approaches relies on preceding user input, this weak assessment can be explained by the “cold start phenomena”: Without any user feedback, the system cannot identify users’ interests. Both groups, however, developed a more positive perception, hence indicating that at later stages of the experiment, the displayed sub categories became more focused. Note, however, that the agreement is rather fluctuant. Moreover, a Wilcoxon rank-sum test did not report a significant difference between both groups for the above introduced differentials.

Further, we asked to judge the following two statements: (1) “The displayed sub categories represent my diverse interests in various topics” and (2) “the displayed results for each sub category were related to each other”. Figures 7.8 and 7.9 show the average answers over the whole time of the experiment.

As can be seen, the average ratings from both groups indicate a positive tendency towards the two statements. Thus, their responses suggest that implicit relevance feedback can be used to capture users’ long-term interests. Note, however, that the agreement is rather fluctuant.

7.4. Results

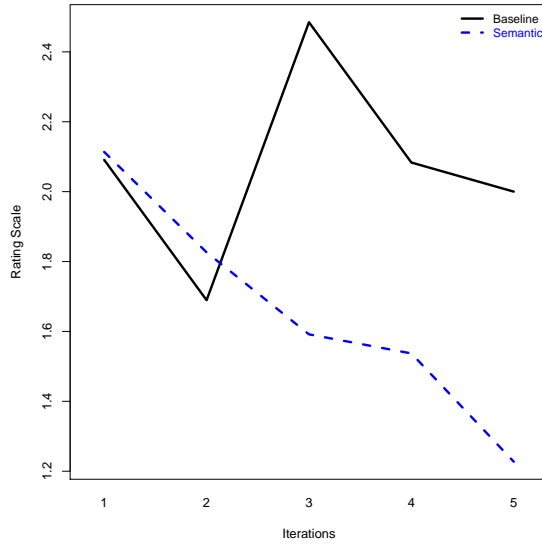


FIGURE 7.8: The displayed sub categories represent my diverse interests in various topics. (lower is better)

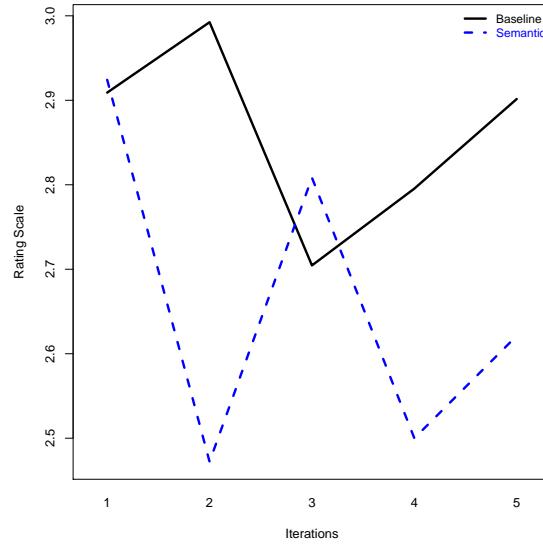


FIGURE 7.9: The displayed results for each sub category were related to each other. (lower is better)

7.4.3 Recommendation Quality

Finally, aiming to evaluate the second hypothesis that ontologies can be used to recommend news videos, we asked the participants to judge whether “the displayed results for each category contained relevant stories I didn’t receive otherwise”. This statement aimed to evaluate whether the recommendations provided some novelty to the user and did not just consist of news documents that the users had already seen before. Furthermore, we asked them to assess the statement “the displayed results for each category matched with the category description”. With this statement, we aimed to understand whether the recommendations were in the right context. The participants’ replies are depicted in Figures 7.10 and 7.11.

Note that even though no significance can be reported, the semantic-based recommendation run received an overall higher weighting than the baseline run. Thus, the results support the outcome of our simulation that ontologies can be successfully employed to provide news video recommendations. Aiming to evaluate this observation, we further analysed the users’ search behaviour. Figures 7.12 and 7.13 show the average number of recommendations that were triggered by the users, i.e. the average number of clicks on the sub categories and the average number of manually triggered searches, respectively.

While users of the semantic recommender system constantly triggered recommendations over all days of the experiment, Figure 7.12 shows a less homogeneous interac-

7.4. Results

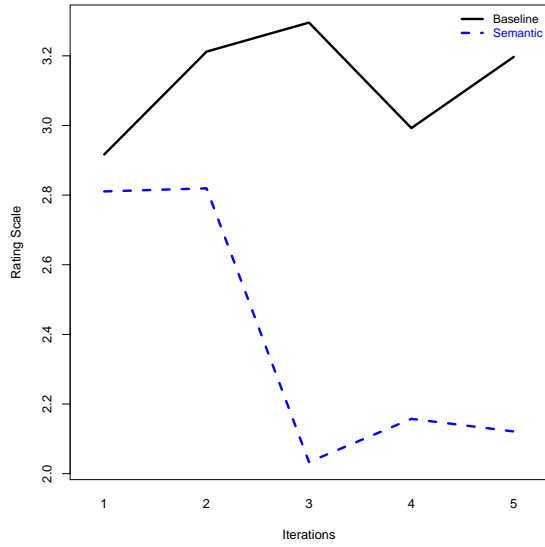


FIGURE 7.10: The displayed results for each category contained relevant stories I didn't receive otherwise. (lower is better)

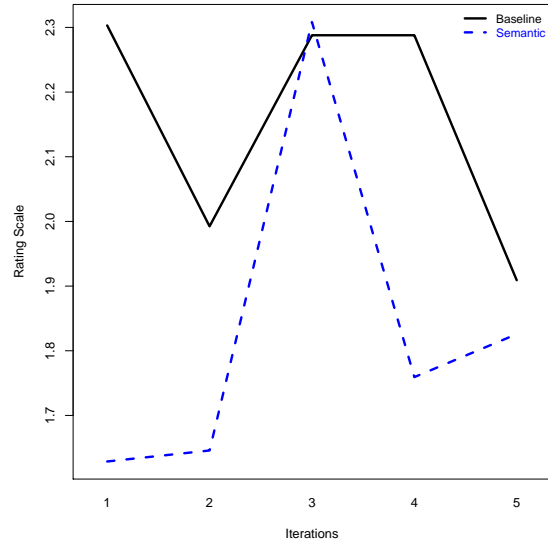


FIGURE 7.11: The displayed results for each category matched with the category description. (lower is better)

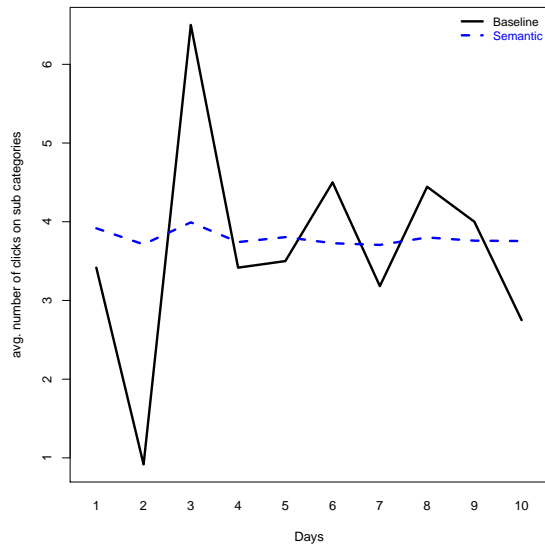


FIGURE 7.12: The average number of clicks on the sub categories

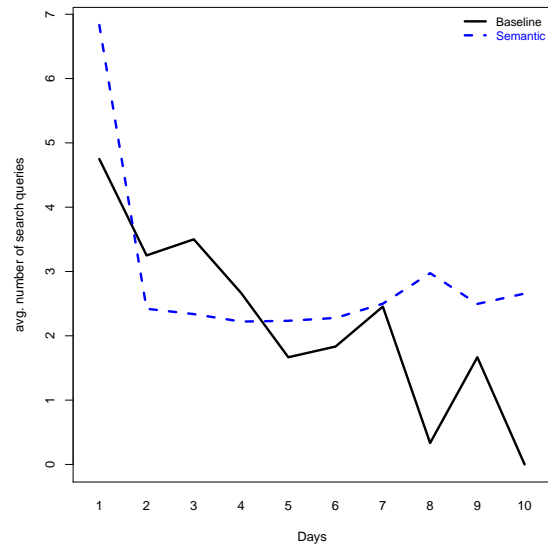


FIGURE 7.13: The average number of manually triggered searches

tion pattern of the baseline system's user group. Moreover, Figure 7.13 shows a decline in the amount of manually triggered search queries of the baseline group, while the second group constantly triggered their own searches. Both observations could indicate the dissatisfaction that the users of the baseline system experienced during the study. The longer they interacted with the system, the less convinced they seemed to be of its ability to retrieve relevant videos. Moreover, the irregular usage of recommendations might indicate that the users interacted with the system rather randomly, maybe just to fulfil the search task.

7.4.4 Discussion

In this section, we evaluated the two hypotheses introduced by analysing both the questionnaires and log files of our user study. As we have shown before, the participants were asked to express their opinion about the recommender system by filling in interim questionnaires at various stages of the user study. We analysed these questionnaires against three different criteria. First, we aimed to evaluate the general usability of the system and the way that users interacted with it. The participants' responses indicate that they used the recommender system to explore various types of news topics and that the system was helpful. This observation can be backed up by a statement that one participant formulated in the exit questionnaire of the experiment: "In general, the system is great to explore news according to the user interests. The automatic organisation of topics and interest evolved through the time I used the system, and I did not need to search again using the keyword box. That was definitely nice." Thus, we conclude that the introduced system has the potential to improve the users' news gathering routines.

Further, we discussed various questions that aimed to evaluate whether implicit relevance feedback can be used to capture users' interest over a longer period. The users' responses suggest that the introduced technique successfully captured users' broad interests and was able to successfully identify sub interests.

Finally, we introduced the users' judgements about the recommendation quality. Even though positive tendencies could be spotted by both user groups, the semantic-based recommender system achieved a better rating than the baseline system. Thus, we conclude that our semantic-based recommendation technique can successfully be employed to provide novel and relevant news recommendations over a longer time period.

7.5 Summary

In this chapter, we aimed to confirm the outcome of the simulation-based evaluation of long-term news video recommendation that has been outlined in Chapter 6 by employing a multi-session time-series user study. We suggested an experimental methodology where 24 users were asked to include an online news video recommender system over a time period of ten days into their daily news gathering routine. The news recommender system automatically captures daily broadcasting news and segments the bulletins into coherent news stories. We evaluated two types of recommendations that have been evaluated and fine-tuned in the previous chapter. Differing from standard interactive information retrieval experiments, this evaluation was split into multiple sessions and was performed under an uncontrolled environment, two necessary conditions for a realistic evaluation of implicit user profiling. This novel approach cannot rely on system-centred evaluation measures as common in information retrieval experiments. Thus, standardised evaluation measures need yet to be developed. The hypotheses were evaluated by analysing users' feedback and interactions which were provided during various stages of the experiment. The analysis appears to support the simulation-based evaluation of the previous chapter. We therefore conclude that both our implicit user profiling methodology and semantic-based recommendation technique can be used to capture users' long-term interests and to recommend relevant video documents.

– *We can only see a short distance ahead, but we can see plenty there that needs to be done.*

Alan Turing, 1950

8

Conclusion and Future Work

This chapter concludes this thesis. First, we summarise the contributions of this work in Section 8.1. Conclusions that can be drawn from these contributions are discussed in Section 8.2. Limitations are discussed in Section 8.3. Future research directions are outlined in Section 8.4.

8.1 Summary

The following contributions have been presented in this thesis:

- Video retrieval interfaces provide various facilities that can be employed to gather evidence of users' interests. Assuming that users interact with documents they find relevant, we identified implicit indicators of relevance by analysing representative video retrieval interfaces and then studied the effect that different weighting schemes for these features can have on the retrieval performance by introducing a simulation-based evaluation scheme.
- We have introduced a short term video recommendation approach that exploits the implicit interactions of other users to provide collaboration-based video recommendations. We evaluated the recommendation technique by performing a user study where users were asked to retrieve as many video shots as possible for four of the hardest TRECVID search topics.

- After analysing a long-term personalisation scenario, we introduced an architecture for user profiling. Conditions and limitations of this architecture were discussed.
- Focusing on this long-term personalisation scenario, we proposed an implicit user profiling technique. Users interactions are interpreted as implicit relevance feedback and stored in personal profiles. We proposed to structure the profiles to better represent users' interests in multiple topics. Further, we proposed to exploit underlying semantics, i.e. by employing a generic ontology to generate personalised video recommendations. The Ostensive Model is used to model users' developing information need. Both simulation-based and user-centred evaluation schemes are employed to evaluate the introduced semantic user profile.
- The simulation-based evaluation required the creation of a personal test collection that can be used to evaluate various recommendation techniques over multiple iterations. We therefore introduced our approach of creating personalised assessment lists that represent users' personal interests in various news topics over a longer time period. Using these assessment lists, users are simulated as interacting with a video recommender system over multiple iterations. The simulations indicate the success of our recommendation technique, and allows the fine tuning of various parameters of the recommendation technique.
- The user-centred evaluation scheme confirmed the outcome of the simulation. Users were asked to include a video recommender system into their daily news gathering routine and to assess the system in frequent online questionnaires. The introduced evaluation methodology can be applied to other longitudinal studies which aim to evaluate personalised retrieval and recommendation techniques.

8.2 Conclusion

After summarising the main contributions of this thesis in the previous section, this section discusses the main conclusions that can be drawn from the work.

8.2.1 Implicit Indicators of Relevance

White [2004] argues that users' interactions during their information seeking process can be interpreted as their implicit way to express interest in the documents' contents. In a Web scenario, for example, printing or bookmarking a web site can be interpreted as implicit indicator of relevance, i.e. the users performed a certain action *because* they

are interested in the corresponding document. We surveyed various research approaches that illustrate that implicit relevance feedback is, by now, considered to be a strong tool for the personalisation of retrieval results in the text domain. The main challenge that all approaches are facing is that it is not clear which indicators can successfully be employed to infer relevance. Kelly [2004], for example, argues that reading time is a questionable behaviour feature which does not necessarily indicate interest. Consequently, most approaches rely on a combination of implicit interactions, referred to as low-level events.

As we have discussed, however, little work has been done on exploiting implicit relevance feedback in the video domain. Video documents are of significantly different nature when compared to text and image documents, the main difference being the added time dimension. Consequently, users interact with video documents differently when compared with their textual counterparts. The first contribution of this thesis is therefore an analysis of implicit relevance feedback in the video domain. We analysed representative video retrieval interfaces and identified the most common low-level events that can be recorded while interacting with these interfaces. Even though we do not identify the importance of these individual events, we argue that a combination of the features can be used to improve users' information seeking task performance. In this thesis, we studied different scenarios where implicit relevance feedback can be employed. A discussion on these scenarios is given in the remainder of this section.

8.2.2 Collaborative Recommendations

In the first scenario, we studied whether a combination of these low-level features can be used to improve session-based retrieval. Derived from this study, one contribution of this work is a model for exploiting community-based usage information for video retrieval, where implicit usage information from past users are exploited in order to provide enhanced assistance in video retrieval tasks. The model is designed to capture the complex interactions of a user with an interactive video retrieval system, including the representation of sequences of user-system interaction during a search session. As the characteristics of the available implicit information are profoundly related with the characteristics of the retrieval system (e.g. the user interface, the offered interactions mechanism, or the possibility judging the relevance of a result without opening it), our study leads to the conclusion that these characteristics have to be considered in order to select the appropriate recommendation approach.

Building upon the model, we introduced a recommendation technique that exploits this usage information pool. Our experimental setup assumes that users are performing

similar tasks, which can be thought of as a rather strong assumption, but one required by the use of a test collection. Task definitions in TRECVideo are rarely related, making it difficult to test our recommendation approaches with partially related tasks. The results indicate the effectiveness of this technique and hence, it suggests that implicit relevance feedback can be employed in the video retrieval domain to recommend videos within single search sessions.

8.2.3 Implicit User Profiling

Another contribution of this thesis is the evaluation of long-term user profiling techniques. We focused on two main problems within this work.

Firstly, we studied how users' interests can be captured implicitly. We therefore proposed a technique of capturing users' implicit relevance feedback in a user profile. Moreover, aiming to exploit these profiles, we separated profiles based on users' diverse interests, i.e. by categorising the content that users interacted with. Our evaluation indicates that implicit relevance feedback can successfully be employed to generate such profiles.

Another problem we addressed in this thesis was the lack of standardised schemes for the evaluation of longitudinal personalisation techniques. We suggested the development of a new test collection used for studying long-term user modelling techniques in video retrieval. We first introduced an approach of generating independent relevance assessment lists. Exploiting these lists, we propose a simulation-based evaluation scheme. We defined unique interaction patterns and identified usage patterns by exploiting a preceding user study. Moreover, we employed both patterns and assessment lists to generate implicit user profiles. Our simulation-based evaluation illustrates how these user profiles can be used to evaluate long-term personalisation approaches. It enables us to evaluate the performance of different adaptation approaches over multiple iterations. Using a classical evaluation scheme, such an evaluation would have been challenging. The main conclusion of this study is that the introduced data collection can be used for the benchmarking of long-term recommendation approaches. Since all results are achieved by employing a simulation, further runs can be performed to fine tune recommendation parameters. Nevertheless, even though simulations can be used for benchmarking, it cannot replace real user studies.

8.2.4 Long-Term Recommendation

The final contribution of this thesis is how the implicit relevance feedback can be exploited to recommend news videos that match users' long-term interests. Within this

context, we addressed three research challenges.

Firstly, we recommended video documents by formulating personalised search queries using the users' implicit user profiles. As we have shown in Section 2.2, creating search queries based on users' previous interaction is a common technique in the text retrieval domain. The studies which have been discussed in this thesis indicate that the same technique can also be employed in the video domain.

Secondly, we propose to model users' evolving interests using the Ostensive Model. As our evaluation suggests, the model can successfully be employed to identify those stories in the users' profile that they recently showed the highest interest in.

Thirdly, we have shown that contextual semantics can be useful to recommend relevant news stories that match the users' interests. We proposed to set news videos into their semantic context by relying on a generic ontology. According to Thomas [2010], a similar idea is followed by DailyMe, Inc., who introduced in August 2009 their personalisation service called "Newstogram". Using similar techniques that have been employed in this thesis, their service can be used to generate personalised news web sites.

8.3 Limitations

After concluding the major contributions of this thesis, this section highlights the limitations of introduced work. The limitations are discussed further in Section 8.4 which outlines future work.

This thesis, as well as most research in the video retrieval domain, is focused on news videos. The reason for this one-sided concentration on news videos is the dominant position of TRECVideo within the research community. The TRECVideo collection allows evaluating different research approaches following the Cranfield paradigm, where users are asked to search for as many video documents as possible for pre-defined search tasks. Cunningham and Nichols [2008], however, have shown that this paradigm does not depict the way people search and find videos. The major difference between the Cranfield paradigm and real life video accessing is that users do not aim to find as many videos as possible for specific tasks, but rather aim for few, but highly relevant videos. As common in the IR community, we limit our study by focusing on standard evaluation metrics that require the users to find as many relevant documents as possible.

Another difference is that very often, users do not need to browse the whole video corpus to find the videos they are interested in. A large amount of videos, so-called viral videos, are found by internet sharing, i.e. through emails and video sharing websites. Consequently, the users' interactions with the video retrieval interface can be

reduced to a minimum, making the identification of their interests using implicit means a challenging task. Besides, the low-level feedback events which have been employed within this thesis as implicit indicators of relevance have been identified by analysing representative video retrieval interfaces. Most of these interfaces have been designed to assist users in exploring homogeneous news video collections, especially the TRECVID collections. Large-scale video collections such as YouTube consist of videos of different origin, topic, length and quality though. It is not premature to assume that different interface features can be required to address this difference. Hence, another limitation of our work is that it is focused on homogeneous data collections that can be explored using state-of-the-art video retrieval interfaces as common within TRECVID.

Another limitation is that our recommendation and personalisation techniques relies on very content rich video material, i.e. videos which are enriched with textual meta-data. Real video collections, e.g. in YouTube, can consist of videos with many textual annotations, whereas other videos have no textual descriptions at all. Applying the same techniques in such inhomogeneous collection would penalise text-less videos.

Finally, we limit our research on topical relevance, as common within TRECVID. As argued before, different relevance types include, amongst others, cognitive relevance. [Cunningham and Nichols \[2008\]](#) have shown that many users explore video collections to find “funny” videos. Such search scenario, however, is not addressed within this thesis. Nonetheless, work which has been presented in this thesis provides a good insight into the role of implicit relevance feedback in the video domain and the application of semantic user profiles.

8.4 Future Work

After discussing the main contributions and limitations in the previous sections, this section highlights future research directions.

8.4.1 Implicit Indicators of Relevance

After identifying potential implicit indicators of relevance in the video retrieval domain, we have shown in this thesis that applying implicit relevance feedback can positively be used to personalise video retrieval in both short-term and long-term adaptation scenarios. However, the range of implicit indicators in a video retrieval application remains unclear. The following research questions should hence be answered: Which actions carried out by a user can be considered as a positive indicator of relevance and can hence be used to adapt retrieval results? The second question is how these features

can be weighted to increase retrieval performance. It is not clear which features are stronger and which are weaker indicators of relevance, respectively. Once the users' intentions and information demand is clear, systems can be built that take advantage of such knowledge and optimise the retrieval output for each user by implementing an adaptive video retrieval model. In [Hopfgartner, 2008], we proposed to study these questions by providing users with different video retrieval interface approaches for different interaction environments, such as desktop PCs and iTV boxes. Users are required to interact differently with the interfaces. The differences may have a strong influence on users' behaviour, making the importance of implicit indicators of relevance application-dependent. Comparing user interactions with different applications should help to identify common positive indicators. The research could be conducted around different applications where user feedback can be monitored, such as desktop computers, television and mobile phones. The specific characteristics of these three environments are introduced in the following.

- *Desktop Computers:* The most familiar environment for the user to do video retrieval is probably a standard desktop computer. Most adaptive video retrieval systems, including the ones in this thesis, have been designed to run under such an environment. The interface can be displayed on the screen, and users can easily interact with the system by using the keyboard or mouse. We can assume that users will take advantage of this interaction and hence give a high quantity of implicit feedback. From today's point of view, this environment offers the highest possibility for implicit relevance feedback.
- *iTV:* A widely accepted medium for multimedia consumption is the television. Watching television, however, is a passive procedure. Viewers can select a program using a remote control, but changing the content is not possible. Recently, interactive TV is becoming increasingly popular. Using a remote control, viewers can interact directly when watching television (e.g. they can participate in quiz shows). In news video retrieval, this limited interaction is a challenge. It will be more complex to enter query terms (e.g. by using the channel selection buttons as is common for remote controls). Hence, users will possibly avoid entering keywords. On the other hand, the selection keys and a display on the remote control provide a method to give explicit relevance feedback. For example, the viewer sees a video segment on television, then uses the remote control to judge the relevance of this segment.
- *Mobile Handhelds:* Within the last decade, a large amount of money and effort went into improving the speed and capability of mobile networks, with the 3GPP

Long Term Evolution (LTE) technology [Ekstrom et al., 2006] being the next standard to come. Simultaneously, mobile devices (“Smartphones”) that are able to handle multimedia content have gained market share. Thus, mobile handhelds are an important environment where users can interact with video content. Due to the rather small size of the device, user interactions might be reduced to a minimum. At the same time, the amount of information that can be displayed on the screen is limited.

The scenario which we focused on is a simplified representation of current practice. More than ever, we are facing a constant information input, resulting in an information overload. For example, we can consume news on the radio, read news headlines while passing a newsagent stand, consume content online or overhear others talking about latest news. All these input can influence our decision to retrieve the corresponding news stories. Modelling these multiple inputs is therefore a challenging task.

8.4.2 Collaborative Recommendations

Another contribution of this thesis is the use of collaborative relevance feedback to recommend related videos. In order to further improve the introduced model, various studies can and have already been performed. An interesting research question is how the graph-based model can be employed to capture users’ long-term interests. The introduced study focuses on satisfying users’ short-term information needs, neglecting long-term interests. An important limiting factor for the study of this question is scalability. In Vallet et al. [2010], we evaluate the scalability of our graph-based technique by introducing a simulation-based evaluation scheme. Overall, the experimental results suggest that collaborative recommendation approaches on video retrieval depend largely on the quality and the amount of the usage information available, and that these factors have to be considered in order to compare different approaches. Further, the scalability experiment did evaluate the effect of noisy data, mostly related to the simulated quality of implicit evidence, but also to partially overlapping tasks. The results suggest that the recommendation performance is similar when similar tasks are used.

Another extension of this work has been introduced by Vrochidis et al. [2010] who enrich our action graph with visual features. Differing from our approach, they consider that users might perform different sub sessions within one session which are not related to each other. Therefore, they create a new interaction pool for each sub session. Actions that are performed within different sub sessions are thus treated separately. At the end of each session, these sub interaction pools are merged. They evaluate their

technique by performing a user study. Similar to our results, they conclude that the implicit interaction pool can be used to improve video search.

A possible method of improving our model is through the use of negative implicit feedback [Yang et al., 2007], which our representation supports. Negative feedback, which takes into account possible negative evidence of a document being irrelevant to the current user's task, has to be handled with care, but could be used in order to achieve a higher quality implicit information pool, and hence better performing recommendation.

8.4.3 Implicit User Profiling

The third main contribution is the creation of implicit user profiles. As we have shown, users' implicit relevance feedback can be used for the identification of users' long-term interests in multiple topics. In Chapter 4, we highlighted that mainstream video retrieval systems such as YouTube and Dailymotion are used by vast numbers of web users and hence argue for the implementation of collaborative relevance feedback. An interesting research question is whether this community-based relevance feedback can also be employed for the creation of implicit user profiles that capture users' long-term interests.

Furthermore, we argued that one of the main challenges in studying user profiling techniques is the evaluation of such approaches. We approached this problem by both employing a simulation-based and user-centred evaluation scheme. Similar to other simulations (e.g. [Nguyen and Worring, 2008; Foley and Smeaton, 2008; Joho et al., 2009; Worring et al., 2007; Vallet et al., 2008b]), the introduced simulation scheme relies on pre-defined interaction patterns, backed by statistical click analyses of a preceding user study. Stereotype users are mimicked by analysing *how often* and under which *conditions* particular events are performed by real users. Hence, the simulation is rather generic and based on heuristic user interactions. Important factors such as the users' motivation are completely ignored. According to Ingwersen and Järvelin [2005], however, user motivation has a strong influence on the information seeking process. It affects how users understand and assess retrieval results, and their ability to manage the retrieval process. User motivation may also impact user articulation capabilities, i.e. their ability to express their information need correctly. A poorly articulated information need tends to include 'hidden', i.e. implicit questions, that the user has not expressed in a clear and explicit way. Indeed, studies (e.g. Ozmutlu and Spink [2001]) have shown that users gradually progress towards shorter queries and fewer results examined. Users become increasingly passive in their browsing behaviour, indicating that

user motivation may gradually become weaker the longer the information seeking process.

Modelling user motivation can help adjust IR systems or their components, according to different user types. Considering the levels of motivation, recommendations can be applied about which system or system component could be selectively applied to different types of users.

This simplified simulation of real users interacting with a retrieval interface allows a preliminary evaluation of underlying research hypotheses. Even though simplicity might be helpful in avoiding disturbing factors which might negatively influence the actual evaluation, it is questionable whether simplified simulations really model a *realistic* user interaction scenario. The main problem of the introduced simulation scheme is that the complex nature of human behaviour, e.g. user motivation, is completely ignored. Statistical log file analyses can reveal how often users interacted with relevant and non-relevant documents, respectively. Following this scheme, the actual reason for their interaction, their motivation, cannot be captured. Motivation is a key factor of human behaviour though. In order to achieve more realistic user simulation models, the users' motivation when interacting with interactive retrieval systems needs to be studied carefully.

Future work should therefore focus on the development of a realistic user interaction simulation which takes different user motivation states into account. Various questions need to be answered in such a realistic user interaction simulation.

1. What is the user's intention when performing a certain action?
2. What motivates users to click on results or to provide feedback?
3. Which external factors (e.g. time of the day, office environment) influence the users in their information seeking process and how should this influence be modelled?
4. Do emotions, triggered by the user's information need, influence the interaction behaviour?
5. How can user actions be predicted?

Answering these questions has the potential to prove a scientific base for simulation-based evaluation of interactive information retrieval systems.

8.4.4 Long-Term Recommendation

Finally, we focused on recommending video documents by exploiting long-term implicit relevance feedback. Following state-of-the-art approaches in the text domain (see Section 2.2), we recommend video documents by creating personalised search queries which are then triggered to retrieve video documents.

One of the main contributions is the use of the Ostensive Model to identify more recent user interests. As we have shown, however, using the Ostensive Model does not support the appearance of “breaking news”, since no representation of this news topic can be found in the users’ profile. Even though a user might be very interested in such breaking news, this interest will be overshadowed by their long-term interests. In order to address this problem, further research should be performed to automatically identify such breaking news and boost them in the recommendations accordingly. One method to identify these events might be to employ Twitter⁸⁻¹. Twitter is a social networking service where users can post short text-based messages online, referred to as “tweeting”. Other users can follow these tweets by subscribing to users’ profiles. Moreover, tweets can be “retweeted”, i.e. forwarded again. Kwak et al. [2010] use Twitter to identify “trending topics”. Analysing millions of tweets, they conclude that over 85% of these topics are news headlines. Further, they argue that tweets get retweeted almost instantly. Consequently, we argue that Twitter offers great potential for the identification of breaking news.

Another contribution is the use of semantics to identify recommendations. As our evaluation indicate, semantic recommendations can successfully be employed to improve the recommendation quality. An open issue is, however, to provide an ontology that contains entities for every concept that appear in news. Within our study, we relied on DBpedia, which is a generic ontology created from an older Wikipedia image. Consequently, latest developments might not be covered by this ontology. Future work should therefore include developing an ontology that is updated as fast as Wikipedia. One solution might be to directly include semantics into Wikipedia. Völkel et al. [2006], for example, propose an extension to be integrated into Wikipedia that eases the generation of a semantic knowledge base. Considering the diverse group of Wikipedia contributors [Ortega, 2009], however, it is questionable whether such extension would result in a constant semantic annotation of Wikipedia content.

⁸⁻¹<http://www.twitter.com/>, last time accessed on: 14 July 2010

Bibliography

- Eytan Adar. User 4XXXXX9: Anonymizing Query Logs. In Einat Amitay, G. Craig Murray, and Jaime Teevan, editors, *Proceedings of the 16th International World Wide Web Conference, Workshop Query Log Analysis: Social and Technological Challenges*, 5 2007.
- John Adcock, Jeremy Pickens, Matthew Cooper, Lisa Anthony, Francine Chen, and Pernilla Qvarfordt. FXPAL Interactive Search Experiments for TRECVID 2007. In Wessel Kraaij, Alan F. Smeaton, and Paul Over, editors, *TRECVID'07: Notebook Papers and Results*, pages 135–144, Gaithersburg, Maryland, USA, 2007. National Institute of Standards and Technology.
- John Adcock, Matthew Cooper, and Jeremy Pickens. Experiments in interactive video search by addition and subtraction. In [Luo et al. \[2008\]](#), pages 465–474. ISBN 978-1-60558-070-8.
- Gediminas Adomavicius and Alexander Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.
- Eugene Agichtein, Eric Brill, and Susan T. Dumais. Improving web search ranking by incorporating user behavior information. In [Efthimiadis et al. \[2006\]](#), pages 19–26. ISBN 1-59593-369-7.
- Philippe Aigrain, HongJiang Zhang, and Dragutin Petkovic. Content-Based Representation and Retrieval of Visual Media: A State-of-the-Art Review. *Multimedia Tools and Applications*, 3(3):179–202, 1996.
- Selim Aksoy and Özge Çavuş. A Relevance Feedback Technique for Multimodal Retrieval of News Videos. In Ljiljana Milic, editor, *EUROCON'05: Proceedings of the International Conference on Computer as a Tool, Belgrade, Serbia & Montenegro*, pages 139–142. IEEE, 11 2005. ISBN 1-4244-0049-X.
- Kamal Ali and Wijnand van Stam. TiVo: making show recommendations using a distributed collaborative filtering architecture. In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors, *KDD'04: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 394–401. ACM, 2004. ISBN 1-58113-888-1.
- Chris Anderson. *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion, 7 2006. ISBN 978-1401302375.
- Joaquim Arlandis, Paul Over, and Wessel Kraaij. Boundary error analysis and categorization in the trecvid news story segmentation task. In [Leow et al. \[2005\]](#), pages 103–112. ISBN 3-540-27858-3.

-
- Farshid Arman, Remi Depommier, Arding Hsu, and Ming-Yee Chiu. Content-based browsing of video sequences. In Meera Blattner and John O. Limb, editors, *MM'94: Proceedings of the Second ACM International Conference on Multimedia, San Francisco, California, USA*, pages 97–103. ACM Press New York, NY, USA, 10 1994.
- Richard Arndt, Raphaël Troncy, Steffen Staab, Lynda Hardman, and Miroslav Vacura. COMM: Designing a Well-Founded Multimedia Ontology for the Web. In Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *ISWC'07: Proceedings of the 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, Busan, Korea*, volume 4825 of *Lecture Notes in Computer Science*, pages 30–43. Springer, 11 2007. ISBN 978-3-540-76297-3.
- Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, Harlow, 1st edition, 1999. ISBN 0-201-39829-X.
- Ricardo A. Baeza-Yates, Nivio Ziviani, Gary Marchionini, Alistair Moffat, and John Tait, editors. *SIGIR'05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil*, 8 2005. ACM. ISBN 1-59593-034-5.
- Werner Bailer, Christian Schober, and Georg Thallinger. Video Content Browsing Based on Iterative Feature Clustering for Rushes Exploitation. In [Over et al. \[2006\]](#), pages 230–239.
- Alex Bailey, Ian Ruthven, and Leif Azzopardi. Evaluating user studies in information access. In Fabio Crestani and Ian Ruthven, editors, *CoLIS'05: Proceedings of the 5th International Conference on Conceptions of Library and Information Sciences, Glasgow, UK*, volume 3507 of *Lecture Notes in Computer Science*, page 251. Springer Verlag, 6 2005. ISBN 978-3-540-26178-0.
- Erwin M. Bakker, Thomas S. Huang, Michael S. Lew, Nicu Sebe, and Xiang Sean Zhou, editors. *CIVR'03: Proceedings of the Second International Conference on Image and Video Retrieval, Urbana-Champaign, IL, USA*, volume 2728 of *Lecture Notes in Computer Science*, 2003. Springer. ISBN 3-540-40634-4.
- Kobus Barnard and David A. Forsyth. Learning the Semantics of Words and Pictures. In Bob Werner, editor, *ICCV'01: Proceedings of the Eighth IEEE International Conference on Computer Vision, Vancouver, British Columbia, Canada*, pages 408–415. IEEE, 2001. ISBN 0-7695-1143-0.
- Travis Bauer and David B. Leake. Real Time User Context Modeling for Information Retrieval Agents. In Henrique Paques, Ling Liu, David Grossman, and Calton Pu, editors, *CIKM'01: Proceedings of the ACM CIKM International Conference on Information and Knowledge Management, Atlanta, Georgia, USA*, pages 568–570. ACM, 2001. ISBN 1-58113-436-3.

-
- BBC. Ceefax marks 30 years of service. http://news.bbc.co.uk/1/hi/entertainment/tv_and_radio/3681174.stm, 9 2004. last time accessed on: 15 April 2010.
- Richard K. Belew and C. J. van Rijsbergen. *Finding out about: a cognitive perspective on search engine technology and the WWW*. Cambridge University Press, New York, NY, USA, 1st edition, 2000. ISBN 0-521-63028-2.
- Richard Kuehn Belew. Adaptive information retrieval: using a connectionist representation to retrieve and learn about documents. *SIGIR Forum*, 23(SI):11–20, 1989. ISSN 0163-5840.
- Nicholas J. Belkin. Some(what) grand challenges for information retrieval. *SIGIR Forum*, 42(1):47–54, 2008.
- Nicholas J. Belkin, Colleen Cool, Diane Kelly, Shin jeng Lin, Soyeon Park, Jose Perez Carballo, and C. Sikora. Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Information Processing and Management*, 37(3):403–434, 2001.
- Allan Bell. *The Language of News Media*. Wiley-Blackwell, 6 1991. ISBN 978-0631164357.
- Tim Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 284(5):34–43, 5 2001.
- Marco Bertini, Alberto Del Bimbo, and Carlo Torniai. Multimedia enriched ontologies for video digital libraries. *International Journal of Parallel, Emergent and Distributed Systems*, 22(6):407–416, 2007.
- Matthias Bezold. Describing user interactions in adaptive interactive systems. In [Houben et al. \[2009\]](#), pages 150–161. ISBN 978-3-642-02246-3.
- J. Bhogal, Andy Macfarlane, and P. Smith. A review of ontology based query expansion. *Information Processing & Management: An International Journal*, 43(4):866–886, 2007. ISSN 0306-4573.
- Daniel Billsus, Clifford Brunk, Craig Evans, Brian Gladish, and Michael J. Pazzani. Adaptive interfaces for ubiquitous web access. *Communications of the ACM*, 45(5): 34–38, 2002.
- David C. Blair. Some thoughts on the reported results of TREC. *Information Processing and Management*, 38(3):445–451, 2002.
- Henk M. Blanken, Arjen P. de Vries, Henk Ernst Bok, and Ling Feng. *Multimedia Retrieval*. Springer Verlag, Heidelberg, Germany, 1st edition, 2007. ISBN 978-3540728948.

-
- David M. Blei and Michael I. Jordan. Modeling annotated data. In [Callan et al. \[2003\]](#), pages 127–134. ISBN 1-58113-646-3.
- Eric Bloedorn, Inderjeet Mani, and T. Richard MacMillan. Machine Learning of User Profiles: Representational Issues. In *AAAI/IAAI, Vol. 1*, pages 433–438. AAAI, 1996.
- Paolo Boldi and Sebastiano Vigna. MG4J at TREC 2006. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006, Gaithersburg, Maryland, November 14-17, 2006*, volume Special Publication 500-272. National Institute of Standards and Technology (NIST), 2006.
- Susanne Boll, Qi Tian, Lei Zhang, Zili Zhang, and Yi-Ping Phoebe Chen, editors. *MMM’10: Proceedings of the 16th International Multimedia Modeling Conference, Chongqing, China*, volume 5916 of *Lecture Notes in Computer Science*, 2010. Springer. ISBN 978-3-642-11300-0.
- Pia Borlund. The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 2003a.
- Pia Borlund. The concept of relevance in IR. *JASIST*, 54(10):913–925, 2003b.
- Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soulé-Dupuy, editors. *ECIR’09: Proceedings of the 31th European Conference on IR Research, Toulouse, France*, volume 5478 of *Lecture Notes in Computer Science*, 4 2009. Springer Verlag. ISBN 978-3-642-00957-0.
- Paul Browne, Csaba Czirik, Georgina Gaughan, Cathal Gurrin, Gareth Jones, Hyowon Lee, Sean Marlow, Kieran Mc Donald, Noel Murphy, Noel O’Connor, Neil O’Hare, Alan F. Smeaton, and Jiamin Ye. Dublin City University Video Track Experiments for TREC 2003. In Alan F. Smeaton, Wessel Kraaij, and Paul Over, editors, *TREC Vid’03: Notebook Papers and Results*, Gaithersburg, Maryland, USA, 2003. National Institute of Standards and Technology.
- Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors. *The Adaptive Web*. Springer Verlag, 2007. ISBN 978-3-540-72078-2.
- John Buford and Scott Stevens, editors. *MM’99: Proceedings of the 7th ACM International Conference on Multimedia, Orlando, Florida, USA, 10 1999*. ACM. ISBN 1-58113-151-8.
- Martin Bulmer. *Questionnaires V.1*. Thousand Oakes, CA: Sage Publications, 2004.
- Tobias Bürger, Erich Gams, and Georg Güntner. Smart content factory: assisting search for digital objects by generic linking concepts to multimedia content. In Siegfried Reich and Manolis Tzagarakis, editors, *HT’05: Proceedings of the 16th ACM Conference on Hypertext and Hypermedia, Salzburg, Austria*, pages 286–287. ACM, 2005. ISBN 1-59593-168-6.

-
- Marius-Gabriel Butuc. Semantically Enriching Content Using OpenCalais. In Stefan-Gheorghe Pentiu, editor, *EDITIA'09: Proceedings of the Romanian Workshop on Distributed Systems, Suceava, Romania*, pages 77–88, 12 2009.
- Jamie Callan, Gordon Cormack, Charles Clarke, David Hawking, and Alan F. Smeaton, editors. *SIGIR'03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada*, 2003. ACM. ISBN 1-58113-646-3.
- Donald T. Campbell and Julian C. Stanley. *Experimental and Quasi-Experimental Design for Research*. Wadsworth Publishing, Monterey, CA, 1st edition, 1963. ISBN 978-0395307878.
- Iain Campbell. Supporting information needs by ostensive definition in an adaptive information space. In Ian Ruthven, editor, *MIRO'95: Proceedings of the Final Workshop on Multimedia Information Retrieval, Glasgow, Scotland, UK*, Workshops in Computing. BCS, 9 1995.
- Iain Campbell. Interactive Evaluation of the Ostensive Model Using a New Test Collection of Images with Multiple Relevance Assessments. *Information Retrieval*, 2(1): 85–112, 2000.
- Iain Campbell and C. J. van Rijsbergen. The ostensive model of developing information needs. In Peter Ingwersen and Niels Ole Pors, editors, *CoLIS'06: Proceedings of the 6th International Conference on Conceptions of Library and Information Sciences, Copenhagen, Denmark*, pages 251–268. The Royal School of Librarianship, 10 1996.
- Murray Campbell, Alexander Haubold, Shahram Ebadollahi, Milind R. Naphade, Apostol Natsev, Joachim Seidl, John R. Smith, Jelena Tešić, and Lexing Xie. IBM Research TRECVID-2006 Video Retrieval System. In [Over et al. \[2006\]](#), pages 175–182.
- David Carmel, Hagai Roitman, and Naama Zwerdling. Enhancing cluster labeling using wikipedia. In James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel, editors, *SIGIR'09: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 139–146. ACM, 2009. ISBN 978-1-60558-483-6.
- Les Carr, David De Roure, Arun Iyengar, Carole A. Goble, and Michael Dahlin, editors. *WWW'06: Proceedings of the 15th International Conference on World Wide Web, Edinburgh, Scotland, UK*, 2006. ACM. ISBN 1-59593-323-9.
- Jason Chaffee and Susan Gauch. Personal Ontologies for Web Navigation. In Arvin Agah, Jamie Callan, Elke Rundensteiner, and Susan Gauch, editors, *CIKM'00: Proceedings of the 2000 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA*, pages 227–234. ACM, 11 2000.

-
- Lekha Chaisorn and Tat-Seng Chua. The segmentation and classification of story boundaries in news video. In Xiaofang Zhou and Pearl Pu, editors, *VDB'02: Proceedings of the IFIP TC2/WG2.6 Sixth Working Conference on Visual Database Systems, Brisbane, Australia*, volume 216 of *IFIP Conference Proceedings*, pages 95–109. Kluwer, 2002. ISBN 1-4020-7060-8.
- Shih-Fu Chang, R. Manmatha, and Tat-Seng Chua. Combining Text and Audio-Visual Features in video Indexing. In Billene Mercer, editor, *ICASSP'05: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1005–1008. IEEE, 3 2005. ISBN 0-7803-8874-7.
- David Wai-Lok Cheung, Il-Yeol Song, Wesley W. Chu, Xiaohua Hu, and Jimmy J. Lin, editors. *CIKM'09: Proceedings of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China, 2009*. ACM. ISBN 978-1-60558-512-3.
- Ed H. Chi, Peter Pirolli, Kim Chen, and James E. Pitkow. Using information scent to model user information needs and actions and the web. In Julie Jacko and Andrew Sears, editors, *CHI'01: Proceedings of the SIG-CHI on Human factors in Computing Systems, Seattle, WA, USA*, pages 490–497. ACM Press, 3 2001.
- Paul-Alexandru Chirita, Wolfgang Nejdl, Raluca Paiu, and Christian Kohlschütter. Using odp metadata to personalize search. In [Baeza-Yates et al. \[2005\]](#), pages 178–185. ISBN 1-59593-034-5.
- Choicestream, Inc. 2008 Choicestream Personalization Survey. Technical Report, Choicestream, Inc., 210 Broadway, Fourth Floor, Cambridge, MA, USA, 2008.
- Michael G. Christel. Establishing the utility of non-text search for news video retrieval with real world users. In [Lienhart et al. \[2007\]](#), pages 707–716. ISBN 978-1-59593-702-5.
- Michael G. Christel. Examining user interactions with video retrieval systems. In *SPIE'06: Proceedings of SPIE Volume 6506, Multimedia Content Access: Algorithms and Systems*. Society of Photo-Optical Instrumentation Engineers, 2007b. ISBN 9780819466198.
- Michael G. Christel. Supporting video library exploratory search: when storyboards are not enough. In [Luo et al. \[2008\]](#), pages 447–456. ISBN 978-1-60558-070-8.
- Michael G. Christel and Ronald M. Conescu. Addressing the challenge of visual information access from digital image and video libraries. In Mary Marlino, Tamara Sumner, and Frank M. Shipman III, editors, *JCDL'05: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, Denver, CA, USA, June 7-11, 2005*, pages 69–78. ACM, 2005. ISBN 1-58113-876-8.
- Vassilis Christophides and Georgia Koutrika, editors. *PersDB'08: Proceedings of the Second International Workshop on Personalized Access, Profile Management, and Context Awareness: Databases*, 2008.

-
- Tat-Seng Chua, Shih-Fu Chang, Lekha Chaisorn, and Winston H. Hsu. Story boundary detection in large broadcast news video archives: techniques, experience and trends. In Henning Schulzrinne, Nevenka Dimitrova, Martina Angela Sasse, Sue B. Moon, and Rainer Lienhart, editors, *ACM MM'04: Proceedings of the 12th ACM International Conference on Multimedia, October 10-16, 2004, New York, NY, USA*, pages 656–659. ACM, 2004. ISBN 1-58113-893-8.
- Mark Claypool, Phong Le, Makoto Wased, and David Brown. Implicit interest indicators. In Candy Sidner and Johanna Moore, editors, *IUI'01: Proceedings of the International Conference on Intelligent User Interfaces, Santa Fe, New Mexico, USA*, pages 33–40. ACM, 1 2001.
- Cyril Cleverdon. The effect of variations in relevance assessments in comparative experimental tests of index languages. Technical report, Cranfield Institute of Technology, 10 1970.
- Cyril Cleverdon, Jack Mills, and Michael Keen. Factors determining the performance of indexing systems. Technical report, ASLIB Cranfield project, Cranfield, 1966.
- Louise Cole. Copyright in the digital age: a UK perspective. *The E-Resources Management Handbook*, 2 2009. Published online at: <http://uksg.metapress.com/openurl.asp?genre=article&id=doi:10.1629/9552448-0-3.15.1>, last time accessed on: 14 May 2010.
- Nick Craswell and Martin Szummer. Random walks on the click graph. In [Kraaij et al. \[2007\]](#), pages 239–246. ISBN 978-1-59593-597-7.
- W. Bruce Croft and David J. Harper. Using probabilistic models of document retrieval without relevance information. *Readings in information retrieval*, pages 339–344, 1997.
- W. Bruce Croft, Robert Cook, and Dean Wilder. Providing Government Information on the Internet: Experiences with THOMAS. In David M. Levy and Richard Furuta, editors, *DL'95: Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries*, pages 19–24, Austin, TX, 6 1995.
- W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors. *SIGIR'98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia, New York, NY, USA, 2007*. ACM Press. ISBN 1-58113-015-5.
- Carlos A. Cuadra. Opening the Black Box of Relevance. *Journal of Documentation*, 42(1):291–303, 1967.
- Sally Jo Cunningham and David M. Nichols. How people find videos. In [Larsen et al. \[2008\]](#), pages 201–210. ISBN 978-1-59593-998-2.
- Mark Czerwinski, Eric Horvitz, and Edward Cutrell. Subjective duration assessment: An implicit probe for Software usability. In *Proceedings of IHM-HCI 2001 Conference, Lille, France*, pages 167–170, 2001.

-
- Mariam Daoud, Lynda Tamine-Lechani, and Mohand Boughanem. Using a graph-based ontological user profile for personalizing search. In James G. Shanahan, Sihem Amer-Yahia, Ioana Manolescu, Yi Zhang, David A. Evans, Aleksander Kolcz, Key-Sun Choi, and Abdur Chowdhury, editors, *CIKM'08: Proceedings of the 17th ACM Conference on Information and Knowledge Management, Napa Valley, California, USA*, pages 1495–1496. ACM, 2008. ISBN 978-1-59593-991-3.
- Mariam Daoud, Lynda Tamine-Lechani, Mohand Boughanem, and Bilal Chebaro. A session based personalized search using an ontological user profile. In Sung Y. Shin and Sascha Ossowski, editors, *SAC'09: Proceedings of the 2009 ACM Symposium on Applied Computing, Honolulu, Hawaii*, pages 1732–1736. ACM, 2009. ISBN 978-1-60558-166-8.
- Abhinandan Das, Mayur Datar, Ashutosh Garg, and ShyamSundar Rajaram. Google News Personalization: Scalable Online Collaborative Filtering. In [Williamson et al. \[2007\]](#), pages 271–280. ISBN 978-1-59593-654-7.
- Ali Dasdan, Kostas Tsioutsoulis, and Emre Velipasaoglu. Web search engine metrics: (direct metrics to measure user satisfaction). In [Rappa et al. \[2010\]](#), pages 1343–1344. ISBN 978-1-60558-799-8.
- H. Day. Looking time as a function of stimulus variables and individual differences. *Perceptual & Motor Skills*, 22(2):423–428, 1966.
- Abdigani Diriye, Srdan Zagorac, Suzanne Little, and Stefan Rüger. NewsRoom: An Information-Seeking Support System for News Videos. In James Wang, Nozha Boujemaa, Nuria Oliver Ramirez, and Apostol Natsev, editors, *MIR'10: Proceedings of the 11th ACM SIGMM International Conference on Multimedia Information Retrieval*, pages 377–380, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-815-5.
- Alan Dix, Janet Finlay, and Russell Beale. Analysis of user behaviour as time series. In *HCI'92: Proceedings of the Conference on People and computers VII*, pages 429–444, New York, NY, USA, 1993. Cambridge University Press. ISBN 0-521-44591-4.
- Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. A large-scale evaluation and analysis of personalized search strategies. In [Williamson et al. \[2007\]](#), pages 581–590. ISBN 978-1-59593-654-7.
- A. Doulamis and N. Doulamis. Performance evaluation of Euclidean/correlation-based relevance feedback algorithms in content-based image retrieval systems. In Luis Torres, editor, *ICIP'03: Proceedings of the International Conference on Image Processing, Barcelona, Spain*, volume 4, pages 737–740. IEEE, 09 2003. ISBN 0-7803-7750-8.
- Anastasios D. Doulamis, Yannis S. Avrithis, Nikolaos D. Doulamis, and Stefanos D. Kollias. Interactive content-based retrieval in video databases using fuzzy classification and relevance feedback. In *ICMCS, Vol. 2*, pages 954–958, 1999.

-
- M. Carl Drott. Using web server logs to improve site design. In Kathy Haramundanis, Laurie Bennett, and Phyllis Galt, editors, *SIGDOC'98: Proceedings of the 16th annual International Conference on Computer Documentation*, pages 43–50, New York, NY, USA, 9 1998. ACM Press.
- Minko Dudev, Shady Elbassuoni, Julia Luxenburger, Maya Ramanath, and Gerhard Weikum. Personalizing the Search for Knowledge. In [Christophides and Koutrika \[2008\]](#), pages 1–8.
- Gwenaël Durand, Gabriella Kazai, Mounia Lalmas, Uwe Rauschenbach, and Patrick Wolf. A Metadata Model Supporting Scalable Interactive TV Services. In Yi-Ping Phoebe Chen, editor, *MMM'05: Proceedings of the 11th International Conference on Multi Media Modeling, Melbourne, Australia*, pages 386–391. IEEE Computer Society, 1 2005. ISBN 0-7695-2164-9.
- Efthimis N. Efthimiadis, Susan T. Dumais, David Hawking, and Kalervo Järvelin, editors. *SIGIR'06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA*, 8 2006. ACM. ISBN 1-59593-369-7.
- Hannes Ekstrom, Anders Furuskar, Jonas Karlsson, Michael Meyer, Stefan Parkvall, Johan Torsner, and Mattias Wahlqvist. Technical solutions for the 3g long-term evolution. *Communications Magazine, IEEE*, 44(3):38–45, 2006.
- Abdulmotaleb El Saddik and Son Vuong, editors. *MM'08: Proceedings of the 16th ACM International Conference on Multimedia, Vancouver, Canada*, 2008. ACM.
- Desmond Elliott and Joemon M. Jose. A proactive personalised retrieval system. In [Cheung et al. \[2009\]](#), pages 1935–1938. ISBN 978-1-60558-512-3.
- Desmond Elliott, Frank Hopfgartner, Teerapong Leelanupab, Yashar Moshfeghi, and Joemon M. Jose. An Architecture for Life-long User Modelling. In Judy Kay and Bob Kummerfeld, editors, *Proceedings of the Lifelong User Modelling Workshop, Trento, Italy*, pages 9–16, 6 2009.
- T. Faw and J. Nunnally. The Effects on Eye Movements of Complexity, Novelty, and Affective Tone. *Perception & Psychophysics*, 2(7):263–267, 1967.
- Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998. ISBN 026206197X.
- Miriam Fernández, Vanessa López, Marta Sabou, Victoria Uren, David Vallet, Enrico Motta, and Pablo Castells. Using TREC for cross-comparison between classic IR and ontology-based search models at a Web scale. In Marko Grobelnik, Peter Mika, Thanh Tran Duc, and Haofen Wang, editors, *SemSearch'09: Workshop on Semantic Search at the 18th International World Wide Web Conference, Madrid, Spain*, 4 2009.
- Colum Foley and Alan F. Smeaton. Evaluation of coordination techniques in synchronous collaborative information retrieval. In Jeremy Pickens, Gene Golovchinsky,

-
- and Meredith Ringel Morris, editors, *Proceedings of the First Collaborative Search Workshop, Pittsburgh, Pennsylvania, USA*, 7 2008.
- Jill Freyne, Rosta Farzan, Peter Brusilovsky, Barry Smyth, and Maurice Coyle. Collecting community wisdom: integrating social search & social navigation. In David N. Chin, Michelle X. Zhou, Tessa A. Lau, and Angel R. Puerta, editors, *IUI'07: Proceedings of the 2007 International Conference on Intelligent User Interfaces, Honolulu, Hawaii, USA*, pages 52–61. ACM, 2007. ISBN 1-59593-481-2.
- G. W. Furnas. Generalized fisheye views. *ACM SIGCHI Bulletin*, 17(4):16–23, 1986.
- Dora C. Gálvez Cruz. *An Environment for Protecting the Privacy of E-Shoppers*. PhD thesis, University of Glasgow, 2009.
- Herbert J. Gans. *Deciding What's News: A Study of CBS Evening News, NBC Nightly News, Newsweek, and Time*. Northwestern University Press, Evanston, Illinois, USA, 25th anv edition, 2 2005. ISBN 978-0810122376.
- Susan Gauch, Mirco Speretta, Aravind Chandramouli, and Alessandro Micarelli. *User Profiles for Personalized Information Access*, chapter 2, pages 54–89. Volume 1 of [Brusilovsky et al. \[2007\]](#), 2007. ISBN 978-3-540-72078-2.
- Gary Geisler, Gary Marchionini, Barbara M. Wildemuth, Anthony Hughes, Meng Yang, Todd Wilkens, and Richard Spinks. Video browsing interfaces for the open video project. In Loren Terveen and Dennis Wixon, editors, *CHI'02: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Extended abstracts on Human factors in computing systems*, pages 514–515, New York, NY, USA, 2002. ACM. ISBN 1-58113-454-1.
- Arnab Ghoshal, Sanjeev Khudanpur, J. Magalhães, Simon Overell, and Stefan Rüger. Imperial College and Johns Hopkins University at TRECVID. In [Over et al. \[2006\]](#), pages 183–192.
- Barney G. Glaser and Anselm L. Strauss. *Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine, 6 1967.
- David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12): 61–70, 1992.
- Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(5-6):907–928, 1995.
- Ling Guan and Hong-Jiang Zhang, editors. *ICME'06: Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, Toronto, Ontario, Canada*, 2006. IEEE.
- Cathal Gurrin. Content-based video retrieval. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 466–473. Springer US, 2009. ISBN 978-0-387-35544-3, 978-0-387-39940-9.

-
- Alan Haggerty, Ryen W. White, and Joemon M. Jose. NewsFlash: Adaptive TV News Delivery on the Web. In Andreas Nürnberger and Marcin Detyniecki, editors, *AMR'03: Proceedings of the First International Workshop on Adaptive Multimedia Retrieval, Hamburg, Germany*, volume 3094 of *Lecture Notes in Computer Science*, pages 72–86. Springer, 2004. ISBN 978-3-540-22163-0.
- Martin Halvey and Mark T. Keane. Analysis of online video search and sharing. In Simon Harper, Helen Ashman, Mark Bernstein, Alexandra I. Cristea, Hugh C. Davis, Paul De Bra, Vicki L. Hanson, and David E. Millard, editors, *HT'07: Proceedings of the 18th ACM Conference on Hypertext and Hypermedia, Manchester, UK*, pages 217–226. ACM, 9 2007. ISBN 978-1-59593-820-6.
- Martin Halvey, P. Punitha, David Hannah, Robert Villa, Frank Hopfgartner, Anuj Goyal, and Joemon M. Jose. Diversity, assortment, dissimilarity, variety: A study of diversity measures using low level features for video retrieval. In [Boughanem et al. \[2009\]](#), pages 126–137. ISBN 978-3-642-00957-0.
- Martin Halvey, David Vallet, David Hannah, and Joemon M. Jose. Vigor: a grouping oriented interface for search and retrieval in video libraries. In Fred Heath, Mary Lynn Rice-Lively, and Richard Furuta, editors, *JCDL'09: Proceedings of the 2009 Joint International Conference on Digital Libraries, Austin, TX, USA*, pages 87–96. ACM, 2009b. ISBN 978-1-60558-322-8.
- Martin Halvey, David Vallet, David Hannah, and Joemon M. Jose. Vigor: a grouping oriented interface for search and retrieval in video libraries. In *JCDL '09: Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 87–96, New York, NY, USA, 2009c. ACM. ISBN 978-1-60558-322-8. doi: <http://doi.acm.org/10.1145/1555400.1555415>.
- Micheline Hancock-Beaulieu and Stephen Walker. An evaluation of automatic query expansion in an online library catalogue. *Journal of Documentation*, 48(4):406–421, 1992. ISSN 0022-0418.
- Alan Hanjalic. Shot-Boundary Detection: Unraveled and Resolved? *IEEE Transactions on Circuits and Systems for Video Technology*, 12(2):90–105, 2 2002.
- Alexander K. Hartmann. *A Practical Guide To Computer Simulation*. World Scientific Publishing Co., Inc., 2009.
- Alex Hauptmann, Mike Christel, R. Concescu, J. Gao, Q. Jin, J.Y. Pan, S.M. Stevens, R. Yan, J. Yang, and Y. Zhang. CMU Informedia's TRECVID 2005 Skirmishes. In [Over et al. \[2005\]](#).
- Alexander G. Hauptmann. Lessons for the future from a decade of informedia video analysis research. In [Leow et al. \[2005\]](#), pages 1–10. ISBN 3-540-27858-3.
- Alexander G. Hauptmann, Wei-Hao Lin, Rong Yan, Jun Yang 0003, and Ming yu Chen. Extreme video retrieval: joint maximization of human and computer performance. In [Nahrstedt et al. \[2006\]](#), pages 385–394. ISBN 1-59593-447-2.

-
- Philip J. Hayes, Laura E. Knecht, and Monica J. Cellio. A news story categorization system. *Readings in information retrieval*, pages 518–526, 1997.
- Xuming He, Richard S. Zemel, and Miguel Á. Carreira-Perpiñán. Multiscale conditional random fields for image labeling. In Frances Titsworth, editor, *CVPR'04: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Volume 2*, pages 695–702. IEEE, 2004.
- Marti Hearst. *Search User Interfaces*. Cambridge University Press, 2009.
- Daniel Heesch, Peter Howarth, J. Magalhães, Alexander May, Marcus Pickering, Alexei Yavlinski, and Stefan Rüger. Video Retrieval using Search and Browsing. In Wessel Kraaij, Alan F. Smeaton, and Paul Over, editors, *TRECVID'04: Notebook Papers and Results*, Gaithersburg, Maryland, USA, 2004. National Institute of Standards and Technology.
- William R. Hersh, Andrew Turpin, Susan Price, Benjamin Chan, Dale Kraemer, Lynetta Sacherek, and Daniel Olson. Do batch and user evaluation give the same results? In Emmanuel Yannakoudakis, Nicholas J. Belkin, Mun-Kew Leong, and Peter Ingwersen, editors, *SIGIR'00: Proceedings of the 23rd ACM International Conference on Research and Development in Information Retrieval, Athens, Greece*, pages 17–24. ACM, 7 2000.
- Michiel Hildebrand, Jacco van Ossenbruggen, and Lynda Hardman. An Analysis of Search-based User Interaction on the Semantic Web. Technical Report INS-E0706, Centrum voor Wiskunde en Informatica, 5 2007.
- Frank Hopfgartner. Studying interaction methodologies in video retrieval. *Proceedings of the VLDB Endowment*, 1(2):1604–1608, 2008.
- Frank Hopfgartner and Joemon M. Jose. Evaluating the implicit feedback models for adaptive video retrieval. In [Wang et al. \[2007a\]](#), pages 323–331. ISBN 978-1-59593-778-0.
- Frank Hopfgartner and Joemon M. Jose. On user modelling for personalised news video recommendation. In [Houben et al. \[2009\]](#), pages 403–408. ISBN 978-3-642-02246-3.
- Frank Hopfgartner and Joemon M. Jose. *Toward an Adaptive Video Retrieval System*, chapter 6, pages 113–135. Advances in Semantic Media Adaptation and Personalization. CRC Press: Boca Raton, Florida, 2 edition, 2 2009b. ISBN 978-1420076646.
- Frank Hopfgartner and Joemon M. Jose. Semantic user modelling for personal news video retrieval. In [Boll et al. \[2010\]](#), pages 336–346. ISBN 978-3-642-11300-0.
- Frank Hopfgartner and Joemon M. Jose. Semantic User Profiling Techniques for Personalised Multimedia Recommendation. *Multimedia Systems*, 16:255–274, 2010b.

-
- Frank Hopfgartner, Jana Urban, Robert Villa, and Joemon M. Jose. Simulated Testing of an Adaptive Multimedia Information Retrieval System. In Jenny Benois-Pineau and Eric Pauwels, editors, *CBMI'07: Proceedings of the Fifth International Workshop on Content-Based Multimedia Indexing, Bordeaux, France*, pages 328–335. IEEE, 06 2007. ISBN 1-4244-1010-X.
- Frank Hopfgartner, David Hannah, Nicholas Gildea, and Joemon M. Jose. Capturing Multiple Interests in News Video Retrieval by Incorporating the Ostensive Model. In [Christophides and Koutrika \[2008\]](#), pages 48–55.
- Frank Hopfgartner, Thierry Urruty, Robert Villa, Nicholas Gildea, and Joemon M. Jose. Exploiting Log Files in Video Retrieval. In [Larsen et al. \[2008\]](#), page 454. ISBN 978-1-59593-998-2.
- Frank Hopfgartner, David Vallet, Martin Halvey, and Joemon M. Jose. Collaborative search trails for video search. In Jeremy Pickens, Gene Golovchinsky, and Meredith Ringel Morris, editors, *CIR'08: Proceedings of the First Workshop on Collaborative Information Retrieval, Pittsburgh, USA*, 2008c.
- Frank Hopfgartner, David Vallet, Martin Halvey, and Joemon M. Jose. Search trails using user feedback to improve video search. In [El Saddik and Vuong \[2008\]](#), pages 339–348.
- Frank Hopfgartner, Thierry Urruty, David Hannah, Desmond Elliott, and Joemon M. Jose. Aspect-based video browsing – a user study. In Ching-Yung Lin and Ingemar Cox, editors, *ICME'09: Proceedings of the IEEE International Conference on Multimedia and Expo, New York, USA*, pages 946–949. IEEE, 6 2009. ISBN 978-1-4244-4291-1.
- Frank Hopfgartner, Reede Ren, Thierry Urruty, and Joemon M. Jose. *Information Organisation Issues in Multimedia Retrieval using Low-Level Features*, chapter 15. Multimedia Semantics: Metadata, Analysis and Interaction. Wiley, 1 edition, 2010a. to appear.
- Frank Hopfgartner, Thierry Urruty, Pablo Bermejo, Robert Villa, and Joemon M. Jose. Simulated Evaluation of Faceted Browsing based on Feature Selection. *Multimedia Tools and Applications*, 47(3):631–662, 2010b.
- Geert-Jan Houben, Gord I. McCalla, Fabio Pianesi, and Massimo Zancanaro, editors. *UMAP'09: Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization, formerly UM and AH, Trento, Italy*, volume 5535 of *Lecture Notes in Computer Science*, 2009. Springer. ISBN 978-3-642-02246-3.
- Rui Hu, Stefan M. Rüger, Dawei Song, Haiming Liu, and Zi Huang. Dissimilarity measures for content-based image retrieval. In Joern Ostermann and Touradj Ebrahimi, editors, *ICME'08: Proceedings of the 2008 IEEE International Conference on Multimedia and Expo, Hannover, Germany*, pages 1365–1368. IEEE, 2008.

-
- Chih-Wei Huang. Automatic Closed Caption Alignment Based on Speech Recognition Transcripts. Technical report, University of Columbia, 2003. ADVENT Technical Report.
- Thomas S. Huang and Xiang Sean Zhou. Image retrieval with relevance feedback: From heuristic weight adjustment to optimal learning methods. In [Pitas et al. \[2001\]](#), pages 2–5. ISBN 0-7803-6726-X.
- Francisco Iacobelli, Larry Birnbaum, and Kristian J. Hammond. Tell me more, not just more of the same. In Charles Rich, Qiang Yang, Marc Cavazza, and Michelle X. Zhou, editors, *IUI'10: Proceedings of the International Conference on Intelligent User Interfaces*, pages 81–90. ACM, 2 2010. ISBN 978-1-60558-515-4.
- Peter Ingwersen and Kalvero Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer Verlag, Heidelberg, Germany, 1st edition, 2005. ISBN 978-1402038501.
- Melody Y. Ivory and Marti A. Hearst. The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys*, 33(4):470–516, 2001.
- Alejandro Jaimes, Mike Christel, Sebastien Gilles, Sarukkai Ramesh, and Wei-Ying Ma. Multimedia Information Retrieval: What is it, and why isn't anyone using it? In Hongjiang Zhang, John Smith, and Qi Tian, editors, *MIR'05: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 3–8, New York, NY, USA, 11 2005. ACM Press.
- Ramesh Jain. EventWeb: Developing a Human-Centered Computing System. *Computer*, 41(2):42–50, 2008. ISSN 0018-9162.
- Anthony Jameson. Adaptive Interfaces and Agents. *The human-computer interaction handbook: evolving technologies and emerging applications*, pages 433–458, 2008.
- Bernard J. Jansen, Abby Goodrum, and Amanda Spink. Searching for multimedia: analysis of audio, video and image web queries. *World Wide Web*, 3(4):249–254, 2000a.
- Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management: an International Journal*, 36(2):207–227, 2000b.
- Bernhard J. Jansen. Search log analysis: What it is what's been done, how to do it. *Library & Information Science Research*, 28:407–432, 2006.
- Kalervo Järvelin. Interactive relevance feedback with graded relevance and sentence extraction: simulated user experiments. In [Cheung et al. \[2009\]](#), pages 2053–2056. ISBN 978-1-60558-512-3.
- Kalervo Järvelin, Jaana Kekäläinen, and Timo Niemi. Expansiontool: Concept-based query expansion and construction. *Information Retrieval*, 4(3-4):231–255, 2001. ISSN 1386-4564.

-
- Thorsten Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Claire Nedellec and Céline Rouveirol, editors, *ECML'98: Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer, 1998. ISBN 3-540-64417-2.
- Thorsten Joachims, Laura A. Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In [Baeza-Yates et al. \[2005\]](#), pages 154–161. ISBN 1-59593-034-5.
- Hideo Joho, Jana Urban, Robert Villa, Joemon M. Jose, and C.J. van Rijsbergen, editors. *AIR'06: Proceedings of the First International Workshop on Adaptive Information Retrieval, Glasgow, United Kingdom*, 10 2006.
- Hideo Joho, D. Birbeck, and Joemon M. Jose. An ostensive browsing and searching on the web. In Bich-Lien Doan, Massimo Melucci, and Joemon M. Jose, editors, *CIR'07: Proceedings of the Second International Workshop on Context-Based Information Retrieval*, pages 81–92, 2007.
- Hideo Joho, David Hannah, and Joemon M. Jose. Revisiting IR Techniques for Collaborative Search Strategies. In [Boughanem et al. \[2009\]](#), pages 66–77. ISBN 978-3-642-00957-0.
- Steve Jones, Sally Jo Cunningham, and Rodger J. McNab. Usage Analysis of a Digital Library. In [Witten et al. \[1998\]](#), pages 293–294.
- P. Suman Karthik and C. V. Jawahar. Analysis of Relevance Feedback in Content Based Image Retrieval. *ICARCV'06: Proceedings of the 9th International Conference on Control, Automation, Robotics and Vision*, 2006.
- Diane Kelly. *Understanding implicit feedback and document preference: A naturalistic user study*. PhD thesis, Rutgers University, 2004.
- Diane Kelly and Jaime Teevan. Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum*, 37(2):18–28, 2003. ISSN 0163-5840.
- Diane Kelly, David J. Harper, and Brian Landau. Questionnaire mode effects in interactive information retrieval experiments. *Information Processing and Management: an International Journal*, 44(1):122–141, 2008.
- Diane Kelly, Susan T. Dumais, and Jan O. Pedersen. Evaluation Challenges and Directions for Information-Seeking Support Systems. *IEEE Computer*, 42(3):60–66, 2009.
- Heikki Keskustalo, Kalervo Järvelin, and Ari Pirkola. Evaluating the effectiveness of relevance feedback based on a user simulation model: effects of a user scenario on cumulated gain value. *Information Retrieval*, 11(3):209–228, 2008.

-
- Heikki Keskustalo, Kalervo Järvelin, Ari Pirkola, Tarun Sharma, and Marianne Lykke. Test collection-based ir evaluation needs extension toward sessions - a case of extremely short queries. In Gary Geunbae Lee, Dawei Song, Chin-Yew Lin, Akiko N. Aizawa, Kazuko Kuriyama, Masaharu Yoshioka, and Tetsuya Sakai, editors, *AIRS'09: Proceedings of the 5th Asia Information Retrieval Symposium, Sapporo, Japan*, volume 5839 of *Lecture Notes in Computer Science*, pages 63–74. Springer Verlag, 10 2009. ISBN 978-3-642-04768-8.
- Georgi Kobilarov, Tom Scott, Yves Raimond, Silver Oliver, Chris Sizemore, Michael Smethurst, Christian Bizer, and Robert Lee. Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Connections. In Lora Aroyo, Paolo Traverso, Fabio Ciravegna, Philipp Cimiano, Tom Heath, Eero Hyvönen, Riichiro Mizoguchi, Eyal Oren, Marta Sabou, and Elena Paslaru Bontas Simperl, editors, *ESWC'09: Proceedings of the 6th European Semantic Web Conference, Heraklion, Crete, Greece*, volume 5554 of *Lecture Notes in Computer Science*, pages 723–737. Springer, 2009. ISBN 978-3-642-02120-6.
- Anita Komlodi and Gary Marchionini. Key frame preview techniques for video browsing. In [Witten et al. \[1998\]](#), pages 118–125.
- Anita Komlodi and Laura Slaughter. Visual video browsing interfaces using key frames. In Clare-Marie Karat and Arnold Lund, editors, *CHI '98: Conference summary on Human factors in computing systems, Los Angeles, California, United States*, pages 337–338, New York, NY, USA, 1998. ACM. ISBN 1-58113-028-7.
- Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando, editors. *SIGIR'07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands*, 2007. ACM. ISBN 978-1-59593-597-7.
- Markus Krötzsch, Denny Vrandečić, and Max Völkel. Semantic MediaWiki. In Isabel F. Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Michael Uschold, and Lora Aroyo, editors, *ISWC'06: Proceedings of the 5th International Semantic Web Conference, Athens, GA, USA*, volume 4273 of *Lecture Notes in Computer Science*, pages 935–942. Springer, 2006. ISBN 3-540-49029-9.
- Sanjiv Kumar and Martial Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In Katsushi Ikeuchi, Olivier Faugeras, and Jitendra Malik, editors, *ICCV'03: 9th IEEE International Conference on Computer Vision, Nice, France*, pages 1150–1159. IEEE Computer Society, 2003. ISBN 0-7695-1950-4.
- Giridhar Kumaran and James Allan. Text classification and named entities for new event detection. In [Sanderson et al. \[2004\]](#), pages 297–304. ISBN 1-58113-881-4.
- Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue B. Moon. What is twitter, a social network or a news media? In [Rappa et al. \[2010\]](#), pages 591–600. ISBN 978-1-60558-799-8.

-
- Inald Lagendijk, Alan Hanjalic, Naeem Ramzan, and Martha Larson. PetaMedia Deliverable D6.1 – Integrative Research Plan. <http://www.petamedia.eu/documents/Deliverables/PetaMediaDeliverable6-1-090531-TUD.pdf>, 2009. Last time accessed on: 04 February 2010.
- Ronald L. Larsen, Andreas Paepcke, José Luis Borbinha, and Mor Naaman, editors. *JCDL'08: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, Pittsburgh, PA, USA*, 2008. ACM. ISBN 978-1-59593-998-2.
- Hyowon Lee, Alan F. Smeaton, Noel E. O'Connor, and Barry Smyth. User evaluation of Físchlár-News: An automatic broadcast news delivery system. *ACM Transactions on Information Systems*, 24(2):145–189, 2006.
- Teerapong Leelanupab, Yue Feng, Vassilios Stathopoulos, and Joemon M. Jose. A Simulated User Study of Image Browsing Using High-Level Classification. In Tat-Seng Chua, Yiannis Kompatsiaris, Bernard Mérialdo, Werner Haas, Georg Thallinger, and Werner Bailer, editors, *SAMT'09: Proceedings of the 4th International Conference on Semantic and Digital Media Technologies, Graz, Austria*, volume 5887 of *Lecture Notes in Computer Science*, pages 3–15. Springer, 2009a. ISBN 978-3-642-10542-5.
- Teerapong Leelanupab, Frank Hopfgartner, and Joemon M. Jose. User centred evaluation of a recommendation based image browsing system. In Bhanu Prasad, Pawan Lingras, and Ashwin Ram, editors, *IICAI'09: Proceedings of the 4th Indian International Conference on Artificial Intelligence, Tumkur, Karnataka, India*, pages 558–573. IICAI, 2009b. ISBN 978-0-9727412-7-9.
- Wee Kheng Leow, Michael S. Lew, Tat-Seng Chua, Wei-Ying Ma, Lekha Chaisorn, and Erwin M. Bakker, editors. *CIVR'05: Proceedings of the 4th International Conference on Image and Video Retrieval, Singapore*, volume 3568 of *Lecture Notes in Computer Science*, 2005. Springer. ISBN 3-540-27858-3.
- Henry Lieberman. Letizia: An Agent That Assists Web Browsing. In *IJCAI'95: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada*, pages 924–929. Morgan Kaufmann, 8 1995.
- Rainer Lienhart. Video OCR: A Survey and Practitioner's Guide. *Video Mining*, pages 155–184, 10 2003.
- Rainer Lienhart, Anand R. Prasad, Alan Hanjalic, Sunghyun Choi, Brian P. Bailey, and Nicu Sebe, editors. *MM'07: Proceedings of the 15th International Conference on Multimedia, Augsburg, Germany*, 2007. ACM. ISBN 978-1-59593-702-5.
- Jimmy J. Lin and Mark D. Smucker. How do users find things with PubMed?: towards automatic utility evaluation with user simulations. In Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong, editors, *SIGIR'08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore*, pages 19–26. ACM, 7 2008. ISBN 978-1-60558-164-4.

-
- Christina Lioma and Iadh Ounis. Examining the Content Load of Part of Speech Blocks for Information Retrieval. In *ACL'06: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia*. The Association for Computer Linguistics, 2006.
- Jingjing Liu and Nicholas J. Belkin. Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. In Stephane Marchand-Maillet and Fabio Crestani, editors, *SIGIR'10: Proceedings of the 33rd ACM International Conference on Research and Development in Information Retrieval, Geneva, Switzerland*. ACM, 2010. to appear.
- Ainhoa Llorente and Stefan M. Rüger. Using second order statistics to enhance automated image annotation. In [Boughanem et al. \[2009\]](#), pages 570–577. ISBN 978-3-642-00957-0.
- Ainhoa Llorente, Simon E. Overell, Haiming Liu 0002, Rui Hu, Adam Rae, Jianhan Zhu, Dawei Song, and Stefan M. Rüger. Exploiting Term Co-occurrence for Enhancing Automated Image Annotation. In Carol Peters, Thomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J. F. Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Vivien Petras, editors, *CLEF'08: Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum, Aarhus, Denmark, Revised Selected Papers*, volume 5706 of *Lecture Notes in Computer Science*, pages 632–639. Springer, 2008. ISBN 978-3-642-04446-5.
- Huan-Bo Luan, Shi-Yong Neo, Hai-Kiat Goh, Yong-Dong Zhang, Shouxun Lin, and Tat-Seng Chua. Segregated feedback with performance-based adaptive sampling for interactive news video retrieval. In [Lienhart et al. \[2007\]](#), pages 293–296. ISBN 978-1-59593-702-5.
- Huan-Bo Luan, Yantao Zheng, Shi-Yong Neo, Yongdong Zhang, Shouxun Lin, and Tat-Seng Chua. Adaptive multiple feedback strategies for interactive video search. In [Luo et al. \[2008\]](#), pages 457–464. ISBN 978-1-60558-070-8.
- Jiebo Luo, Ling Guan, Alan Hanjalic, Mohan S. Kankanhalli, and Ivan Lee, editors. *CIVR'08: Proceedings of the 7th ACM International Conference on Image and Video Retrieval, Niagara Falls, Canada*, 2008. ACM. ISBN 978-1-60558-070-8.
- João Magalhães and Stefan M. Rüger. *Semantic Multimedia Information Analysis for Retrieval Applications*, pages 334–354. IDEA group publishing, 2006.
- Paul P. Maglio, Rob Barrett, Christopher S. Campbell, and Ted Selker. Suitor: an attentive information system. In Doug Riecken, David Benyon, and Henry Lieberman, editors, *IUI'00: Proceedings of the Fifth International Conference on Intelligent User Interfaces, Santa Fe, New Mexico, USA*, pages 169–176, New York, NY, USA, 1 2000. ACM Press.
- B. S. Manjunath, Philipe Salembier, and Thomas Sikora, editors. *Introduction to MPEG 7: Multimedia Content Description Language*. Wiley, 6 2002.

-
- Melvin E. Maron and John L. Kuhns. On Relevance, Probabilistic Indexing and Information Retrieval. *Journal of the ACM*, 7(3):216–244, 1960. ISSN 0004-5411.
- Marissa Mayer. Interview at the “LeWeb 2008” Conference in Paris, France, 2008.
- MESH. Multimedia sEmantic Syndication for enHanced news services – ... A view to the future of news. http://www.mesh-ip.eu/upload/MESH_Scenarios.pdf, 2006. Last time accessed on: 25 February 2010.
- Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability (Communications and Control Engineering)*. Springer Verlag, 1996. ISBN 978-3540198321.
- Hemant Misra, Frank Hopfgartner, Anuj Goyal, P. Punitha, and Joemon M. Jose. Tv news story segmentation based on semantic coherence and content similarity. In [Boll et al.](#) [2010], pages 347–357. ISBN 978-3-642-11300-0.
- Mandar Mitra, Amit Singhal, and Chris Buckley. Improving automatic query expansion. In [Croft et al.](#) [2007], pages 206–214. ISBN 1-58113-015-5.
- Bamshad Mobasher. *Data Mining for Web Personalization*, chapter 3, pages 90–135. Volume 1 of [Brusilovsky et al.](#) [2007], 2007. ISBN 978-3-540-72078-2.
- Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8):142–151, 2000.
- James Monaco. *How to Read a Film*. Oxford Press, London, 4th edition, 2009. ISBN 978-0-19-532105-0.
- Masahiro Morita and Yoichi Shinoda. Information filtering based on user behaviour analysis and best match text retrieval. In W. Bruce Croft and C. J. van Rijsbergen, editors, *SIGIR’94: Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, (Special Issue of the SIGIR Forum)*, pages 272–281. ACM/Springer, 1994. ISBN 3-540-19889-X.
- Dan Morris, Meredith Ringel Morris, and Gina Venolia. Searchbar: a search-centric web history for task resumption and information re-finding. In Mary Czerwinski, Arnold M. Lund, and Desney S. Tan, editors, *CHI’08: Proceedings of the 2008 Conference on Human Factors in Computing Systems, Florence, Italy*, pages 1207–1216. ACM, 2008. ISBN 978-1-60558-011-1.
- Steven Morrison and Joemon Jose. A comparative study of online news retrieval and presentation strategies. In Bob Werner, editor, *ISMSE ’04: Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering, Miami, Florida*, pages 403–409, Washington, DC, USA, 12 2004. IEEE Computer Society. ISBN 0-7695-2217-3.

-
- Alexandros Moukas and Pattie Maes. Amalthaea: An Evolving Multi-Agent Information Filtering and Discovery System for the WWW. *Autonomous Agents and Multi-Agent Systems*, 1(1):59–88, 1998. ISSN 1387-2532.
- Javed Mustafa. Seeking Better Web Searches. *Scientific American*, 292:66–73, 2005.
- Klara Nahrstedt, Matthew Turk, Yong Rui, Wolfgang Klas, and Ketan Mayer-Patel, editors. *MM'06: Proceedings of the 14th ACM International Conference on Multimedia*, Santa Barbara, CA, USA, 2006. ACM. ISBN 1-59593-447-2.
- Milind R. Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston H. Hsu, Lyndon S. Kennedy, Alexander G. Hauptmann, and Jon Curtis. Large-Scale Concept Ontology for Multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.
- Shi-Yong Neo, Jin Zhao, Min-Yen Kan, and Tat-Seng Chua. Video Retrieval Using High Level Features: Exploiting Query Matching and Confidence-Based Weighting. In Hari Sundaram, Milind R. Naphade, John R. Smith, and Yong Rui, editors, *CIVR'06: Proceedings of the 5th International Conference on Image and Video Retrieval*, Tempe, AZ, USA, volume 4071 of *Lecture Notes in Computer Science*, pages 143–152. Springer, 2006. ISBN 3-540-36018-2.
- Giang P. Nguyen and Marcel Worring. Interactive Access to Large Image Collections using Similarity-based Visualization. *Journal of Visual Languages and Computing*, 19(2):203–224, 4 2008.
- David M. Nichols. Implicit rating and filtering. In Laszlo Kovacs, editor, *Proceedings of 5th DELOS Workshop on Filtering and Collaborative Filtering*, Budapest, Hungary, pages 31–36. ERCIM, 11 1998.
- Noel E. O'Connor, Csaba Czirik, Seán Deasy, Noel Murphy, Seán Marlow, and Alan F. Smeaton. News story segmentation in the Físchlár video indexing system. In [Pitas et al. \[2001\]](#), pages 418–421. ISBN 0-7803-6726-X.
- A. Oostendorp and D. E. Berlyne. Dimensions in the Perception of Architecture II: Measures of Exploratory Behaviour. *Scandinavian Journal of Psychology*, 19(1): 83–89, 1978.
- Felipe Ortega. *Wikipedia: A Quantitative Analysis*. PhD thesis, Universidad Rey Juan Carlos, Madrid, Spain, 2009.
- Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. Terrier Information Retrieval Platform. In David E. Losada and Juan M. Fernández-Luna, editors, *ECIR'05: Proceedings of the 27th European Conference on IR Research Advances in Information Retrieval*, Santiago de Compostela, Spain, volume 3408 of *Lecture Notes in Computer Science*, pages 517–519. Springer, 2005. ISBN 3-540-25295-9.
- Paul Over, Tzveta Ianeva, Wessel Kraaij, and Alan F. Smeaton, editors. *TRECVID'05: Notebook Papers and Results*, Gaithersburg, Maryland, USA, 2005. National Institute of Standards and Technology.

-
- Paul Over, Wessel Kraaij, and Alan F. Smeaton, editors. *TRECVID'06: Notebook Papers and Results*, Gaithersburg, Maryland, USA, 11 2006. National Institute of Standards and Technology.
- Huseyin Cenk Ozmutlu and Amanda Spink. Time-based analysis of search data logs. In Peter Graham, Muthucumaru Maheswaran, and M. Rasit Eskicioglu, editors, *IC'01: Proceedings of the 2001 International Conference on Internet Computing, Las Vegas, Nevada, USA*, pages 41–46. CSREA Press, 6 2001. ISBN 18925128X.
- Raluca Paiu, Ling Chen, Claudiu S. Firan, and Wolfgang Nejdl. PHAROS - Personalizing Users' Experience in Audio-Visual Online Spaces. In [Christophides and Koutrika \[2008\]](#), pages 40–47.
- Michael J. Pazzani and Daniel Billsus. *Content-Based Recommender Systems*, chapter 10, pages 325–341. Volume 1 of [Brusilovsky et al. \[2007\]](#), 2007. ISBN 978-3-540-72078-2.
- Michael J. Pazzani, Jack Muramatsu, and Daniel Billsus. Syskill & webert: Identifying interesting web sites. In Dan Weld and Bill Clancey, editors, *AAAI/IAAI'96: Proceedings of the 13th National Conference on Artificial Intelligence, Portland, Oregon, USA*, pages 54–61. AAAI, 8 1996.
- Thomas A. Peters. The history and development of transaction log analysis. *Library Hi Tech*, 42(11):41–66, 1993.
- Marcus J. Pickering, Lawrence W. C. Wong, and Stefan M. Rüger. ANSES: Summarisation of News Video. In [Bakker et al. \[2003\]](#), pages 425–434. ISBN 3-540-40634-4.
- Ioannis Pitas, Venetsanopoulos, and Thrasos Pappas, editors. *ICIP'01: Proceedings of International Conference on Image Processing, Thessaloniki, Greece, 2001*. IEEE. ISBN 0-7803-6726-X.
- Dulce Ponceleon, Savitha Srinivasan, Arnon Amir, Dragutin Petkovic, and Dan Diklic. Key to effective video retrieval: effective cataloging and browsing. In Clare-Marie Karat and Arnold Lund, editors, *MM'98: Proceedings of the Sixth ACM international conference on Multimedia, Bristol, United Kingdom*, pages 99–107, New York, NY, USA, 1998. ACM. ISBN 0-201-30990-4.
- Kriengkrai Porkaew and Kaushik Chakrabarti. Query refinement for multimedia similarity retrieval in MARS. In [Buford and Stevens \[1999\]](#), pages 235–238. ISBN 1-58113-151-8.
- Ioannis Psarras and Joemon M. Jose. A system for adaptive information retrieval. In Vincent P. Wade, Helen Ashman, and Barry Smyth, editors, *AH'06: Proceedings of the 4th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Dublin, Ireland*, volume 4018 of *Lecture Notes in Computer Science*, pages 313–317. Springer, 2006. ISBN 3-540-34696-1.

-
- Ioannis Psarras and Joemon M. Jose. Evaluating a personal information assistant. In David Evans, Sadaoki Furui, and Chantal Soulé-Dupuy, editors, *RIAO'07: Proceedings of the 8th International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications)*, Pittsburgh, PA, USA. CID, 2007.
- Guo-Jun Qi, Xian-Sheng Hua, Yong Rui, Jinhui Tang, Tao Mei, and Hong-Jiang Zhang. Correlative multi-label video annotation. In [Lienhart et al. \[2007\]](#), pages 17–26. ISBN 978-1-59593-702-5.
- Michael Rappa, Paul Jones, Juliana Freire, and Soumen Chakrabarti, editors. *WWW'10: Proceedings of the 19th International Conference on World Wide Web, Raleigh, North Carolina, USA*, 2010. ACM. ISBN 978-1-60558-799-8.
- Mika Rautiainen, Timo Ojala, and Tapio Seppänen. Cluster-temporal browsing of large news video databases. In D. T. Lee, editor, *ICME'04: Proceedings of the 2004 IEEE International Conference on Multimedia and Expo, Taipei, Taiwan*, pages 751–754. IEEE, 2004.
- Mika Rautiainen, Matti Varanka, Ilkka Hanski, Matti Hosio, Anu Pramila, Jialin Liu, Timo Ojala, and Tapio Seppänen. TRECVID 2005 Experiments at MediaTeam Oulu. In [Over et al. \[2005\]](#).
- Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In John B. Smith, F. Don Smith, and Thomas W. Malone, editors, *CSCW '94: Proceedings of the 1994 ACM conference on Computer supported cooperative work, Chapel Hill, North Carolina, USA*, pages 175–186. ACM, 1994. ISBN 0-89791-689-1.
- Ronald E. Rice and Christine L. Borgman. The use of computer-monitored data in information science. *Journal of the American Society for Information Science*, 44: 247–256, 1983.
- Ray Richardson, Alan F. Smeaton, and J. Murphy. Using wordnet as a knowledge base for measuring semantic similarity between words. In *AICS'94: Proceedings of the Seventh Annual Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Ireland*, 1994.
- W. L. Richman, S. Kiesler, S. Weisband, and F. Drasgow. A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, 84(5):754–775, 1999.
- Stephen E. Robertson and Micheline Hancock-Beaulieu. On the evaluation of ir systems. *Information Processing and Management: an International Journal*, 28(4): 457–466, 1992.
- Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In *TREC'04: Proceedings of the Text REtrieval Conference, Gaithersburg, USA*, pages 0–, 1994.

-
- Stephen E. Robertson, Steve Walker, and Micheline Beaulieu. Laboratory experiments with Okapi: Participation in the TREC programme. *Journal of Documentation*, 1 (53):20–37, 1997.
- J. J. Rocchio. Relevance feedback in information retrieval. In Gerard Salton, editor, *The SMART retrieval system: experiments in automatic document processing*, pages 313–323, Englewood Cliffs, USA, 1971. Prentice-Hall.
- Yong Rui, Thomas S. Huang, and Sharad Methotra. Relevance Feedback Techniques in Interactive Content-Based Image Retrieval. In *Proceedings of the Sixth Conference on Storage and Retrieval for Image and Video Databases, San Jose, California, USA*, volume 3312, pages 25–36. SPIE, 1998a. ISBN 0-8194-2752-7.
- Yong Rui, Thomas S. Huang, Michael Ortega, and Sharad Mehrotra. Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, 1998b.
- Ian Ruthven. Re-examining the potential effectiveness of interactive query expansion. In [Callan et al. \[2003\]](#), pages 213–220. ISBN 1-58113-646-3.
- Ian Ruthven and Mounia Lalmas. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review*, 18(2):95–145, 2003. ISSN 0269-8889.
- Hidekazu Sakagami and Tomonari Kamba. Learning personal preferences on online newspaper articles from user behaviors. *Computer Networks*, 29(8-13):1447–1455, 1997.
- Gerald Salton, editor. *Automatic text processing: The Transformation Analysis and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1988. ISBN 978-0201122275.
- Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Readings in information retrieval*, pages 355–364, 1997.
- Gerard Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975. ISSN 0001-0782.
- Gerard Salton, Edward A. Fox, and Harry Wu. Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036, 1983. ISSN 0001-0782.
- Gerard Salton, James Allan, and Chris Buckley. Approaches to Passage Retrieval in Full Text Information Systems. In Robert Korfhage, Edie M. Rasmussen, and Peter Willett, editors, *SIGIR'93: Proceedings of the 16th Annual International ACM Conference on Research and Development in Information Retrieval. Pittsburgh, PA, USA, June 27 - July 1, 1993*, pages 49–58. ACM, 1993. ISBN 0-89791-605-0.
- Mark Sanderson. Test Collections for all (Position Paper). In [Joho et al. \[2006\]](#), page 5.

-
- Mark Sanderson and Hideo Joho. Forming test collections with no system pooling. In [Sanderson et al. \[2004\]](#), pages 33–40. ISBN 1-58113-881-4.
- Mark Sanderson, Kalervo Järvelin, James Allan, and Peter Bruza, editors. *SIGIR'04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, 2004*. ACM. ISBN 1-58113-881-4.
- Tefko Saracevic. Relevance reconsidered. In Peter Ingwersen and Niels Ole Pors, editors, *CoLIS'96: Proceedings of the Second International Conference on Conceptions in Library and Information Science, Copenhagen, Denmark*, pages 201–218. Royal School of Librarianship, 10 1996.
- J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. *Collaborative Filtering Recommender Systems*, chapter 9, pages 291–324. Volume 1 of [Brusilovsky et al. \[2007\]](#), 2007. ISBN 978-3-540-72078-2.
- Simon Schenk, Carsten Saathoff, Anton Baumesberger, Frederik Jochum, Alexander Kleinen, Steffen Staab, and Ansgar Scherp. SemaPlorer - Interactive Semantic Exploration of Data and Media based on a Federated Cloud Infrastructure. In *Billion Triples Challenge at the 7th International Semantic Web Conference 2008*, 2008.
- Ansgar Scherp, Thomas Franz, Carsten Saathoff, and Steffen Staab. F—a model of events based on the foundational ontology dolce+DnS ultralight. In Yolanda Gil and Natasha Fridman Noy, editors, *K-CAP'09: Proceedings of the 5th International Conference on Knowledge Capture Redondo Beach, California, USA*, pages 137–144. ACM, 2009. ISBN 978-1-60558-658-8.
- Klaus Schöffmann, Frank Hopfgartner, Oge Marques, Laszlo Böszörményi, and Joe-mon M. Jose. Video browsing interfaces and applications: a review. *SPIE Reviews*, 1(1):018004–1–018004–35, 2010. doi: 10.1117/6.00000005.
- Nicu Sebe and Qi Tian. Personalized multimedia retrieval: the new trend? In [Wang et al. \[2007a\]](#), pages 299–306. ISBN 978-1-59593-778-0.
- Young-Woo Seo and Byoung-Tak Zhang. Learning user’s preferences by analyzing web-browsing behaviors. In *AGENTS'00: Proceedings of the Fourth International Conference on Autonomous Agents, Barcelona, Catalonia, Spain*, pages 381–387, New York, NY, USA, 2000. ACM Press.
- Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. The Semantic Web Revisited. *IEEE Intelligent Systems*, 21(3):96–101, 2006.
- Upendra Shardanand and Pattie Maes. Social information filtering: Algorithms for automating word of mouth. In Irvin R. Katz, Robert Mack, Linn Marks, Mary Beth Rosson, and Jakob Nielsen, editors, *CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems, Denver, Colorado, USA*, pages 210–217, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co. ISBN 0-201-84705-1.

-
- Ryan Shaw, Raphaël Troncy, and Lynda Hardman. Lode: Linking open descriptions of events. In Asunción Gómez-Pérez, Yong Yu, and Ying Ding, editors, *ASWC'09: Proceedings of the Fourth Asian Conference on Semantic Web, Shanghai, China*, volume 5926 of *Lecture Notes in Computer Science*, pages 153–167. Springer, 2009. ISBN 978-3-642-10870-9.
- Xuehua Shen, Bin Tan, and ChengXiang Zhai. Context-sensitive information retrieval using implicit feedback. In [Baeza-Yates et al. \[2005\]](#), pages 43–50. ISBN 1-59593-034-5.
- Ben Shneiderman and Catherine Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In Enrico Bertini, Catherine Plaisant, and Giuseppe Santucci, editors, *BELIV'06: Proceedings of the AVI Workshop on BEyond time and errors: novel evaluation methods for information visualization, Venice, Italy*, pages 1–7. ACM Press, 2006. ISBN 1-59593-562-2.
- Ahu Sieg, Bamshad Mobasher, and Robin D. Burke. Web search personalization with ontological user profiles. In Mário J. Silva, Alberto H. F. Laender, Ricardo A. Baeza-Yates, Deborah L. McGuinness, Bjørn Olstad, Øystein Haug Olsen, and André O. Falcão, editors, *CIKM'07: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, Lisbon, Portugal*, pages 525–534. ACM, 2007. ISBN 978-1-59593-803-9.
- Herbert A. Simon. *Administrative Behavior. A study of decision-making processes in administrative organization*. Macmillan, New York, 1947.
- Nikos Simou, Carsten Saathoff, Stamatia Dasiopoulou, Vaggelis Spyrou, N. Voisine, Vassilis Tzouvaras, Yiannis Kompatsiaris, Yannis Avrithis, and Steffen Staab. An Ontology Infrastructure for Multimedia Reasoning. In Luigi Atzori, Daniele D. Giusto, Riccardo Leonardi, and Fernando Pereira, editors, *VLBV'05: Proceedings of the 9th International Workshop on Visual Content Processing and Representation, Sardinia, Italy*, volume 3893 of *Lecture Notes in Computer Science*. Springer, 9 2005. ISBN 3-540-33578-1.
- Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and TRECVID. In James Ze Wang, Nozha Boujemaa, and Yixin Chen, editors, *MIR'06: Proceedings of the 8th ACM SIGMM International Workshop on Multimedia Information Retrieval, Santa Barbara, California, USA*, pages 321–330. ACM, 10 2006. ISBN 1-59593-495-2.
- Alan F. Smeaton, Paul Over, and Aiden R. Doherty. Video shot boundary detection: seven years of TRECVID activity. *Computer Vision and Image Understanding*, 4 (114):411–418, 2010.
- Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

-
- Mark D. Smucker and James Allan. Find-similar: similarity browsing as a search tool. In [Efthimiadis et al. \[2006\]](#), pages 461–468. ISBN 1-59593-369-7.
- Barry Smyth, Evelyn Balfe, Jill Freyne, Peter Briggs, Maurice Coyle, and Oisín Boydell. Exploiting query repetition and regularity in an adaptive community-based web search engine. *User Modeling and User-Adapted Interaction*, 14(5):383–423, 2004.
- Cees G. Snoek and Marcel Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 2(4):215–322, 2009.
- Cees G. M. Snoek and Marcel Worring. Multimodal Video Indexing: A Review of the State-of-the-art. *Multimedia Tools & Applications*, 25(1):5–35, 2005. ISSN 1380-7501.
- Cees G. M. Snoek, Marcel Worring, Dennis C. Koelma, and Arnold W. M. Smeulders. A Learned Lexicon-Driven Paradigm for Interactive Video Retrieval. *IEEE Transactions on Multimedia*, 9(2):280–292, 2 2007.
- C.G.M. Snoek, K.E.A. van de Sande, O. de Rooij, B. Huurnink, J.C. van Gemert, J.R.R. Uijlings, J. He, X. Li, I. Everts, V. Nedovic, M. van Liempt, R. van Balen, F. Yan, M.A. Tahir, K. Mikolajczyk, J. Kittler, M. de Rijke, J.M. Geusebroek, Th. Gevers, M. Worring, A.W.M. Smeulders, and D.C. Koelma. The mediamill trecvid 2008 semantic video search engine. In Paul Over, Wessel Kraaij, and Alan F. Smeaon, editors, *TRECVID’08: Notebook Papers and Results*, pages 314–327, Gaithersburg, Maryland, USA, 11 2008. National Institute of Standards and Technology.
- Ian Soboroff. Overview of the trec 2004 novelty track. In [Voorhees and Buckland \[2004\]](#).
- Kai Song, YongHong Tian, Wen Gao, and Tiejun Huang. Diversifying the image retrieval results. In [Nahrstedt et al. \[2006\]](#), pages 707–710. ISBN 1-59593-447-2.
- Karen Spärck-Jones. Reflections on TREC. *Information Processing and Management: an International Journal*, 31(3):291–314, 1995.
- Karen Spärck-Jones. Further reflections on TREC. *Information Processing and Management: an International Journal*, 36(1):37–85, 2000.
- Karen Spärck-Jones and C. J. van Rijsbergen. Report on the need for and provision of an ideal information retrieval test collection. Technical report, Computing Laboratory, University of Cambridge, UK, 1965.
- Karen Sparck-Jones, Steve Walker, and Stephen E. Robertson. A probabilistic model of information retrieval: development and comparative experiments - part 1. *Information Processing and Management*, 36(6):779–808, 2000.
- Amanda Spink, Howard Greisdorf, and Judy Bateman. From highly relevant to not relevant: examining different regions of relevance. *Information Processing & Management: An International Journal*, 34(5):599–621, 1998.

-
- Savitha Srinivasan, Dulce B. Poncelson, Arnon Amir, and Dragutin Petkovic. “what is in that video anyway?” in search of better browsing. In *ICMCS’99: Proceedings of the IEEE Intl. Conf. Multimedia and Expo, Vol. 1*, pages 388–393, 1999.
- Vassilios Stathopoulos and Joemon M. Jose. Bayesian mixture hierarchies for automatic image annotation. In [Boughanem et al. \[2009\]](#), pages 138–149. ISBN 978-3-642-00957-0.
- Besiki Stvilia, Les Gasser, Michael B. Twidale, and Linda C. Smith. A Framework for Information Quality Assessment. *Journal of the American Society for Information Science and Technology*, 58(12):1720–1733, 2007.
- Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In [Williamson et al. \[2007\]](#), pages 697–706. ISBN 978-1-59593-654-7.
- S. Sull, J.R. Kim, Y. Kim, H.S. Chang, and S.U. Lee. Scalable hierarchical video summary and search. *Proceedings of SPIE Conference on Storage and Retrieval for Media Databases*, 4315, 2001.
- Jian-Tao Sun, Hua-Jun Zeng, Huan Liu, Yuchang Lu, and Zheng Chen. CubeSVD: a novel approach to personalized Web search. In Allan Ellis and Tatsuya Hagino, editors, *WWW’05: Proceedings of the 14th international conference on World Wide Web, Chiba, Japan*, pages 382–390. ACM, 2005. ISBN 1-59593-046-9.
- H. Tansley. *The multimedia thesaurus: Adding a semantic layer to multimedia information*. PhD thesis, University of Southampton, 2000.
- Francisco Tanudjaja and Lik Mui. Persona: A Contextualized and Personalized Web Search. *HICSS’02: Proceedings of the 35th Hawaii International Conference on System Sciences, Big Island, Hawaii, USA*, 3:67, 2002.
- Jaime Teevan, Susan T. Dumais, and Eric Horvitz. Potential for personalization. *ACM Transactions on Computer-Human Interaction*, 17(1):1–31, 2010. ISSN 1073-0516.
- Krista Thomas. Presentation at the OpenCalais Workshop at WeMedia 2010: Optimizing Content with Semantic Tech. <http://www.slideshare.net/KristaThomas/open-calais-workshop-at-wemedia-2010>, last time accessed on: 22 May 2010, 2010.
- Anastasios Tombros and Mark Sanderson. Advantages of query biased summaries in information retrieval. In [Croft et al. \[2007\]](#), pages 2–10. ISBN 1-58113-015-5.
- Roger Tourangeau, Lance J. Rips, and Kenneth Rasinski. *The Psychology of Survey Response*. Cambridge University Press, 2000.
- Chrisa Tsinaraki, Panagiotis Polydoros, Fotis G. Kazasis, and Stavros Christodoulakis. Ontology-Based Semantic Indexing for MPEG-7 and TV-Anytime Audiovisual Content. *Multimedia Tools and Applications*, 26(3):299–325, 2005.

-
- Alan M. Turing. Computing Machinery and Intelligence. *Mind*, 59(236):433–460, 1950.
- Jana Urban. *An Adaptive Approach for Image Organisation and Retrieval*. PhD thesis, University of Glasgow, 2007.
- Jana Urban, Xavier Hilaire, Frank Hopfgartner, Robert Villa, Joemon M. Jose, Siripinyo Chantamunee, and Yoshihiko Gotoh. Glasgow University at TRECVID 2006. In [Over et al. \[2006\]](#), pages 363–367.
- Jana Urban, Joemon M. Jose, and C. J. van Rijsbergen. An adaptive technique for content-based image retrieval. *Multimedia Tools & Applications*, 31(1):1–28, 2006b.
- David Vallet, Frank Hopfgartner, Martin Halvey, and Joemon M. Jose. Community based feedback techniques to improve video search. *Signal, Image and Video Processing*, 2(4):289–306, 2008a.
- David Vallet, Frank Hopfgartner, and Joemon M. Jose. Use of implicit graph for recommending relevant videos: A simulated evaluation. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White, editors, *ECIR’08: Proceedings of the 30th European Conference on IR Research, ECIR 2008, Glasgow, UK*, volume 4956 of *Lecture Notes in Computer Science*, pages 199–210. Springer, 2008b. ISBN 978-3-540-78645-0.
- David Vallet, Frank Hopfgartner, Joemon M. Jose, and Pablo Castells. Effects of Usage based Feedback on Video Retrieval: A Simulation based Study. *ACM Transactions on Information Systems*, 2010.
- C. J. van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann Ltd, London, 2nd edition, 1979. ISBN 978-0408709293.
- Roelof van Zwol, Lluís García Pueyo, Georgina Ramírez, Börkur Sigurbjörnsson, and M Labad. Video Tag Game. In *WWW’08: Proceedings of the 17th International World Wide Web Conference (WWW developer track)*. ACM Press, 2008.
- Roelof van Zwol, Vanessa Murdock, Lluís García Pueyo, and Georgina Ramírez. Diversifying image search with user generated content. In Michael S. Lew, Alberto Del Bimbo, and Erwin M. Bakker, editors, *MIR’08: Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval, Vancouver, British Columbia, Canada*, pages 67–74. ACM, 2008. ISBN 978-1-60558-312-9.
- Robert Villa, Nicholas Gildea, and Joemon M. Jose. A study of awareness in multimedia search. In [Larsen et al. \[2008\]](#), pages 221–230. ISBN 978-1-59593-998-2.
- Robert Villa, Nicholas Gildea, and Joemon M. Jose. FacetBrowser: A User Interface for Complex Search Tasks. In [El Saddik and Vuong \[2008\]](#).
- Max Völkel, Markus Krötzsch, Denny Vrandečić, Heiko Haller, and Rudi Studer. Semantic wikipedia. In [Carr et al. \[2006\]](#), pages 585–594. ISBN 1-59593-323-9.

-
- Timo Volkmer and Apostol Natsev. Exploring automatic query refinement for text-based video retrieval. In [Guan and Zhang \[2006\]](#), pages 765–768.
- Johann Wolfgang von Goethe. *Die Wahlverwandtschaften*. Cotta, Tübingen, 1 edition, 1809.
- Ellen Voorhees. Building Test Collections for Adaptive Information Retrieval: What to Abstract for What Cost? (Invited Talk). In [Joho et al. \[2006\]](#), page 5.
- Ellen M. Voorhees. The philosophy of information retrieval evaluation. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *CLEF'01: Revised Papers of the Second Workshop of the Cross-Language Evaluation Forum, Evaluation of Cross-Language Information Retrieval Systems*, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370. Springer, 2001. ISBN 3-540-44042-9.
- Ellen M. Voorhees. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, Massachusetts, USA, 2005.
- Ellen M. Voorhees. On test collections for adaptive information retrieval. *Information Processing & Management: An International Journal*, 44(6):1879–1885, 2008.
- Ellen M. Voorhees and Lori P. Buckland, editors. *TREC'04: Proceedings of the Thirteenth Text REtrieval Conference, Gaithersburg, Maryland*, volume Special Publication 500-261, 2004. National Institute of Standards and Technology (NIST).
- Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, MA, USA, 1st edition, 2005. ISBN 978-0262220736.
- Stefanos Vrochidis, Ioannis Kompatsiaris, and Ioannis Patras. Optimizing Visual Search with Implicit User Feedback in Interactive Video Retrieval. In Shipeng Li, editor, *CIVR'10: Proceedings of the 9th ACM International Conference on Image and Video Retrieval, Xi'an, China*. ACM, 2010. to appear.
- James Ze Wang, Nozha Boujemaa, Alberto Del Bimbo, and Jia Li, editors. *MIR'07: Proceedings of the 9th ACM SIGMM International Workshop on Multimedia Information Retrieval, Augsburg, Bavaria, Germany*, 2007a. ACM. ISBN 978-1-59593-778-0.
- Meng Wang, Xian-Sheng Hua, Xun Yuan, Yan Song, and Li-Rong Dai. Optimizing multi-graph learning: towards a unified video annotation scheme. In [Lienhart et al. \[2007\]](#), pages 862–871. ISBN 978-1-59593-702-5.
- P. Wang, M. Berry, and Y. Yang. Mining longitudinal web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology*, 54:743–758, 2003.
- Doug Warner, Stephen D. Durbin, J. Neal Richter, and Zuzana Gedeon. Adaptive web sites: user studies and simulation. In [Carr et al. \[2006\]](#), pages 975–976. ISBN 1-59593-323-9.

-
- John Watkinson. *The MPEG Handbook*. Focal Press, 2 edition, 2001.
- Paul Watzlawick, Janet H. Beavin, and Don D. Jackson. *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies, and Paradoxes*. Norton, New York, NY, 1967. ISBN 9780393010091.
- Stuart Weibel. The Dublin Core: A Simple Content Description Model for Electronic Resources. *Bulletin of the American Society for Information Science and Technology*, 24(1):9–11, 2005.
- Diana Weiß, Johannes Scheuerer, Michael Wenleder, Alexander Erk, Mark Gülbahar, and Claudia Linnhoff-Popien. A user profile-based personalization system for digital multimedia content. In Sofia Tsekeridou, Adrian David Cheok, Konstantinos Giannakis, and John Karigiannis, editors, *DIMEA'08: Proceedings of the Third International Conference on Digital Interactive Media in Entertainment and Arts, Athens, Greece*, volume 349 of *ACM International Conference Proceeding Series*, pages 281–288. ACM, 2008. ISBN 978-1-60558-248-1.
- Alan Wexelblat and Pattie Maes. Footprints: History-rich tools for information foraging. In Marian G. Williams and Mark W. Altom, editors, *CHI '99: Proceedings of the SIGCHI Conference on Human factors in computing systems, Pittsburgh, Pennsylvania, USA*, pages 270–277. ACM, 1999. ISBN 0-201-48559-1.
- Ryen White. *Implicit Feedback for Interactive Information Retrieval*. PhD thesis, University of Glasgow, 2004.
- Ryen White, Ian Ruthven, and Joemon M. Jose. Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In Kalvero Järvelin, Micheline Beaulieu, Ricardo Baeza-Yates, and Sung Hyon Myaeng, editors, *SIGIR'02: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland*, pages 57–64. ACM, 2002.
- Ryen W. White. Using searcher simulations to redesign a polyrepresentative implicit feedback interface. *Information Processing and Management: an International Journal*, 42(5):1185–1202, 2006.
- Ryen W. White, Joemon M. Jose, and Ian Ruthven. A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Information Processing & Management: An International Journal*, 39(5):707–733, 2003. ISSN 0306-4573.
- Ryen W. White, Ian Ruthven, Joemon M. Jose, and C. J. van Rijsbergen. Evaluating implicit feedback models using searcher simulations. *ACM Transactions on Information Systems*, 23(3):325–361, 2005.
- Ryen W. White, Mikhail Bilenko, and Silviu Cucerzan. Studying the use of popular destinations to enhance web search interaction. In [Kraaij et al. \[2007\]](#), pages 159–166. ISBN 978-1-59593-597-7.

-
- Lynn Wilcox, Shingo Uchihashi, Jonathan Foote, Andreas Girgensohn, and John Boreczky. Video Manga: generating semantically meaningful video summaries. In [Buford and Stevens \[1999\]](#), pages 383–392. ISBN 1-58113-151-8.
- Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1 (6):80–83, 1945.
- Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy, editors. *WWW'07: Proceedings of the 16th International Conference on World Wide Web, Banff, Alberta, Canada, 2007*. ACM. ISBN 978-1-59593-654-7.
- Max L. Wilson, m. c. schraefel, and Ryen W. White. Evaluating advanced search interfaces using established information-seeking models. *Journal of the American Society for Information Science and Technology*, 60(7):1407–1422, 2009.
- Ian Witten, Rob Akscyn, and Frank M. Shipman, editors. *DL'98: Proceedings of the 3rd ACM International Conference on Digital Libraries, Pittsburgh, PA, USA, 3 1998*. ACM.
- Marcel Worring, Ork de Rooij, and Ton van Rijn. Browsing visual collections using graphs. In [Wang et al. \[2007a\]](#), pages 307–312. ISBN 978-1-59593-778-0.
- Rong Yan and Alexander G. Hauptmann. A review of text and image retrieval approaches for broadcast news video. *Information Retrieval*, 10(4-5):445–484, 2007.
- Rong Yan, Alexander G. Hauptmann, and Rong Jin. Multimedia search with pseudo-relevance feedback. In [Bakker et al. \[2003\]](#), pages 238–247. ISBN 3-540-40634-4.
- Bo Yang, Tao Mei, Xian-Sheng Hua, Linjun Yang, Shi-Qiang Yang, and Mingjing Li. Online Video Recommendation Based on Multimodal Fusion and Relevance Feedback. In Nicu Sebe and Marcel Worring, editors, *CIVR'07: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, Amsterdam, The Netherlands*, pages 73–80. ACM, 2007. ISBN 978-1-59593-733-9.
- M.M. Yeung and Boon-Lock Yeo. Video visualization for compact representation and fast browsing of pictorial content. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(5):771–785, 1997.
- Rui Yong, Thomas S. Huang, Sharad Mehrotra, and Michael Ortega. A relevance feedback architecture for content-based multimedia information systems. In Penny Storms, editor, *Proceedings of IEEE Workshop on Content-Based Access of Image and Video Libraries, San Juan, Puerto Rico, 1997*.
- Hugo Zaragoza, Nick Craswell, Michael J. Taylor, Suchi Saria, and Stephen E. Robertson. Microsoft Cambridge at TREC 13: Web and Hard Tracks. In [Voorhees and Buckland \[2004\]](#).
- Bernhard P. Zeigler. *Theory of Modelling and Simulation*. Krieger Publishing Co., Inc., 1984.

-
- Yun Zhai, Jingen Liu, and Mubarak Shah. Automatic Query Expansion for News Video Retrieval. In [Guan and Zhang \[2006\]](#), pages 965–968.
- Benyu Zhang, Hua Li, Yi Liu, Lei Ji, Wensi Xi, Weiguo Fan, Zheng Chen, and Wei-Ying Ma. Improving web search results using affinity graph. In [Baeza-Yates et al. \[2005\]](#), pages 504–511. ISBN 1-59593-034-5.
- Hong Jiang Zhang, Jianhua Wu, Di Zhong, and Stephen W. Smoliar. An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, 30(4): 643–658, 1997.
- Xiang Sean Zhou and Thomas S. Huang. Exploring the Nature and Variants of Relevance Feedback. In Lorretta Palagi, editor, *CBAIVL '01: Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries, Fort Collins, Colorado, USA*, pages 94–100, Washington, DC, USA, 6 2001. IEEE Computer Society. ISBN 0-7695-0034-X.
- Xiang Sean Zhou and Thomas S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 8(6):536–544, 2003.
- Justin Zobel and Alistair Moffat. Inverted files for text search engines. *ACM Computer Surveys*, 38(2):215–231, 2006.



Exploiting Implicit Relevance Feedback: Experimental Documents

This appendix presents the experimental documents described in Section 4.4. These include:

- A.1: Information Sheet
- A.2: Consent Form
- A.3: Entry Questionnaire
- A.4: Post-Search Questionnaire
- A.5: Exit Questionnaire

INFORMATION SHEET

Project: A Study of Using Implicit Feedback Techniques to Improve Video Search

Researcher: Frank Hopfgartner, David Vallet, Martin Halvey



**UNIVERSITY
of
GLASGOW**

You are invited to take part in a research study. Before you decide to do so, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully. Ask me if anything is not clear or if you would like more information.

The aim of this experiment is to investigate the relative effectiveness of two different multimedia search systems. We cannot determine the value of search systems unless we ask those people who are likely to be using them, which is why we need to run experiments like these. Please remember that it is the systems, not you, that are being evaluated.

It is up to you to decide whether or not to take part. If you decide to take part you will be given this information sheet to keep and asked to sign a consent form. You are free to withdraw at any time without giving a reason. You also have the right to withdraw retrospectively any consent given, and to require that any data gathered on you be destroyed.

The experiment will last around two hours and you will receive a reward of £10 upon completion. You will be given a chance to learn how to use the two systems before we begin. At this time you will also be asked to complete an introductory questionnaire. You will perform four tasks in total. Each task should take between 15 minutes to complete. After using each system you will be asked to fill in a questionnaire and your interactions (e.g. mouse clicks and key presses) will also be logged. You are encouraged to comment on each interface as you use it, which I will take notes on. Please ask questions if you need to and please let me know about your experience during the search. Finally, after completing all tasks, you will be asked some questions about the tasks, your search strategy and the systems. Remember, you can opt out at any time during the experiment. You will still be rewarded for your effort depending on the number of tasks completed.

All information collected about you during the course of this study will be kept strictly confidential. You will be identified by an ID number and all information about you will have your name and contact details removed so that you cannot be recognised from it. Data will be stored for analysis, and then destroyed.

The results of this study may be used for some PhD research. The results are likely to be published in early 2008. You can request a summary of the results in the consent form. You will not be identified in any report or publication that arises from this work.

This study is being funded by the European Semedia and K-Space projects at the Department of Computer Science, University of Glasgow. This project has been reviewed by the Faculty of Information and Mathematical Sciences Ethics Committee.

For further information about this study please contact:

Frank Hopfgartner
Department of Computing Science, University of Glasgow
17 Lilybank Gardens
Glasgow, G12 8QQ
Email: hopfgarf@dc.s.gla.ac.uk
Tel.: 0141 330 2998

CONSENT FORM



**UNIVERSITY
of
GLASGOW**

Project: **A Study of Using Implicit Feedback
Techniques to Improve Video Search**

Researcher: **Frank Hopfgartner, David Vallet, Martin
Halvey**

Please tick box

1. I confirm I have read and understand the information sheet for the above study and have had the opportunity to ask questions. ☐
2. I understand that my permission is voluntary and that I am free to withdraw at any time, without giving any reason, without my legal rights being affected. ☐
3. I agree to take part in the above study. ☐
4. I would like to receive a summary sheet of the experimental findings ☐

If you wish a summary, please leave an email address

Name of Participant

Date

Signature

Researcher

Date

Signature

ENTRY QUESTIONNAIRE

This questionnaire will provide us with background information that will help us analyse the answers you give in later stages of this experiment. You are not obliged to answer a question, if you feel it is too personal.



**UNIVERSITY
of
GLASGOW**

User ID:

Please place a TICK ☒ in the square that best matches your opinion.

Part 1: PERSONAL DETAILS

This information is kept completely confidential and no information is stored on computer media that could identify you as a person.

1. Please provide your AGE:

2. Please indicate your GENDER:

Male..... ☐ 1

Female..... ☐ 2

3. Please provide your current OCCUPATION:

YEAR:

4. What is your FIELD of work or study?

5. What is your educational level

Undergraduate/No Degree..... ☐ 1

Graduate Student/Primary Degree. ☐ 2

Researcher/Advanced Degree..... ☐ 3

Faculty/Research Staff..... ☐ 4

6. How would you describe your proficiency with ENGLISH

Native Speaker..... ☐ 1

Advanced..... ☐ 2

Intermediate..... ☐ 3

Beginner..... ☐ 4

Part 2: SEARCH EXPERIENCE

Experience with Multimedia

Circle the number closest to your experience.

How often do you...	Never	Once or twice a year	Once or twice a month	Once or twice a week	Once or twice a day	More often
7. deal with videos, photographs or images in your work, study or spare time?	1	2	3	4	5	6
8. take videos or photographs in your work, study or spare time?	1	2	3	4	5	6
9. carry out image or video searches at home or work?	1	2	3	4	5	6
10. follow news stories/events?	1	2	3	4	5	6
11. watch news videos online?	1	2	3	4	5	6

Multimedia Search Experience

12. Please indicate which online search services you use to search for MULTIMEDIA (mark AS MANY as apply)

- Google (<http://www.google.com>)..... ☐ 1
- Yahoo (<http://www.yahoo.com>)..... ☐ 2
- AltaVista (<http://www.altavista.com>)..... ☐ 3
- AlltheWeb (<http://www.alltheweb.com>)..... ☐ 4
- YouTube (<http://www.youtube.com>)..... ☐ 5
- Flickr (<http://www.flickr.com>)..... ☐ 6
- Microsoft (<http://www.live.com>)..... ☐ 7

Others (please specify).....

13. Using the MULTIMEDIA search services you chose in question 12 is GENERALLY:

- | | | | | | | |
|------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|-------------|
| easy | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | difficult |
| stressful | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | relaxing |
| simple | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | complex |
| satisfying | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | frustrating |

N/A

☐

A.3. Entry Questionnaire

14. You find what you are searching for on any kind of MULTIMEDIA search service...

Never					Always					N/A	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	2	3	4	5							

15. Please indicate which systems you use to MANAGE your MULTIMEDIA (mark AS MANY as apply)

None (I just create directories and files on my computer).....	<input type="checkbox"/>	1
Adobe Album.....	<input type="checkbox"/>	2
Picasa (Google).....	<input type="checkbox"/>	3
iView Multimedia (Mac).....	<input type="checkbox"/>	4
ACDSee.....	<input type="checkbox"/>	5
Others (please specify).....	<input type="text"/>	

16. Using the multimedia management tools you chose in question 15 is GENERALLY:

<div style="display: flex; justify-content: space-between; width: 100%;"> easy difficult </div> <div style="display: flex; justify-content: space-between; width: 100%;"> stressful relaxing </div> <div style="display: flex; justify-content: space-between; width: 100%;"> simple complex </div> <div style="display: flex; justify-content: space-between; width: 100%;"> satisfying frustrating </div>										N/A	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

17. It is easy to find a particular image that you have saved previously on your computer...

Never					Always					N/A	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	2	3	4	5							

18. Describe your natural search strategy either online or on your computer (taking a typical search task into consideration)? (Optional)

- Your problem solving strategy?
- Is it dependent on the type of media you are seeking?
- In an ideal scenario, how could a system support your search strategy?

POST-SEARCH QUESTIONNAIRE

To evaluate the system you have just used, we now ask you to answer some questions about it. Take into account that we are interested in knowing your opinion: answer questions freely, and consider there are no right or wrong answers. Please remember that we are evaluating the system you have just used and not you.



UNIVERSITY
of
GLASGOW

User ID:		System:		Task:	
----------	--	---------	--	-------	--

Please place a TICK ☒ in the square that best matches your opinion. Please answer all questions.

Part 1: TASK

In this section we ask about the search tasks you have just attempted.

1.1. The task we asked you to perform were:

unclear	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	clear
easy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	difficult
simple	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	complex
unfamiliar	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	familiar

1.2. It was easy to formulate initial queries on these topics.

<div> <div>Agree</div> <div>Disagree</div> </div>				
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1


1.3. The search I have just performed was.

stressful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	relaxing
interesting	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	boring
tiring	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	restful
easy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	difficult

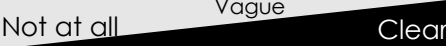
Part 2: RETRIEVED VIDEOS

In this section we ask you about the videos you found/selected.


2.1. The videos I have received through the searches were:

		
relevant	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	not relevant
inappropriate	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	appropriate
complete	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	incomplete
surprising	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	expected


2.2. I had an idea of which kind of videos were relevant for the topic before starting the search.

		
Not at all	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Clear
1	2	3
4	5	


2.3. During the search I have discovered more aspects of the topic than initially anticipated.

		
Disagree	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Agree
1	2	3
4	5	


2.4. The video(s) I chose in the end match what I had in mind before starting the search.

		
Exactly	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Not at all
5	4	3
2	1	

2.5. I believe I have seen all possible videos that satisfy my requirement.

		
Agree	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Disagree
5	4	3
2	1	

2.6. My idea of what videos and terms were relevant changed through out the task.

		
Agree	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Disagree
5	4	3
2	1	

A.4. Post-Search Questionnaire

2.7. I am satisfied with my search results.

Very				Not at all
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

Part 3: SYSTEM & INTERACTION

In this section we ask you some general questions about the system you have just used.

3.1. Overall reaction to the system:

terrible	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	wonderful
satisfying	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	frustrating
easy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	difficult
effective	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	ineffective

3.2. When interacting with the system, I felt:

in control	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	not in control
uncomfortable	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	comfortable
confident	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	unconfident

3.3. How easy was it to USE the system?

Extremely				Not at all
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

3.4. Did you find that the system response time was fast enough?

Extremely				Not at all
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

Part 4: SYSTEM SUPPORT

In this section we ask you more detailed questions about the system and your search strategy.

4.1. The system was effective for solving the task.

<div style="display: flex; justify-content: space-between; width: 100%;"> Agree Disagree </div>				
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

Because it helped me to...

<div style="display: flex; justify-content: space-between; width: 100%;"> Disagree Agree </div>				
---	--	--	--	--

4.2. explore the collection.

1	2	3	4	5
---	---	---	---	---

4.3. find relevant videos.

1	2	3	4	5
---	---	---	---	---

4.4. detect and express different aspects of the task.

1	2	3	4	5
---	---	---	---	---

4.5. focus my search.

1	2	3	4	5
---	---	---	---	---

4.6. find videos that I would not have otherwise considered.

1	2	3	4	5
---	---	---	---	---

4.7. How you conveyed relevance to the system (i.e. marking as relevant, irrelevant) was:

	<div style="display: flex; justify-content: space-around; width: 100%;"> difficult effective not useful </div>					
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
						<div style="display: flex; justify-content: space-around; width: 100%;"> easy ineffective useful </div>

4.8. Do you have any other comments on the system? (optional)

e.g. a) Did selecting images as relevant usually improve the results?
b) What could be improved?

4.9. I believe I have succeeded in my performance of the task.

<div style="display: flex; justify-content: space-between; width: 100%;"> Disagree Agree </div>				
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	2	3	4	5

What are the issues/problems that affected your performance?

<div style="display: flex; justify-content: space-between; width: 100%;"> Agree Disagree </div>				
---	--	--	--	--

4.10. I didn't understand the task.

1	2	3	4	5
---	---	---	---	---

4.11. The video collection didn't contain the video(s) I wanted.

1	2	3	4	5
---	---	---	---	---

4.12. The system didn't return relevant videos.

1	2	3	4	5
---	---	---	---	---

4.13. I didn't have enough time to do an effective search.

1	2	3	4	5
---	---	---	---	---

4.14. I was often unsure of what action to take next.

1	2	3	4	5
---	---	---	---	---

EXIT QUESTIONNAIRE/INTERVIEW

The aim of this experiment was to investigate the relative effectiveness of two different video search systems. Please consider the entire search experience that you just had when you respond to the following questions.



UNIVERSITY
of
GLASGOW

User ID:

Please place a TICK ☒ in the square that best matches your opinion. Please answer the questions as fully as you feel able to.

Which of the systems did you...	System 1	System 2	No difference
1 ... find BEST overall?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2 ... find easier to LEARN TO USE?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3 ... find easier to USE?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4 ... PERFER?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5 ... find changed your perception of the task?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6 ... find more EFFECTIVE for the tasks you performed?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

7 What did you LIKE about each of the systems?

System 1:

System 2:

A.5. Exit Questionnaire

8 What did you DISLIKE about each of the systems?

System 1:

System 2:

9 Additional Comments (Optional)

B

Generating Personalised Ground Truth Data: Experimental Documents

This appendix presents the experimental documents described in Section 6.4. These include:

- B.1: Information Sheet
- B.2: Technical Information Sheet
- B.3: Consent Form
- B.4: Entry Questionnaire
- B.5: Post-Task Questionnaire
- B.6: Exit Questionnaire

Note that the participants were asked to fill in questionnaires (B.4 – B.6) online.

INFORMATION SHEET

Project: Creating personalised ground truth data

Researcher: Frank Hopfgartner



**UNIVERSITY
of
GLASGOW**

You are invited to take part in a research study. Before you decide to do so, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully. Ask me if anything is not clear or if you would like more information.

The aim of this experiment is to identify interesting stories in a news data collection. Please remember that we are not judging your personal interests but rather use your feedback as a base for various user simulations.

It is up to you to decide whether or not to take part. If you decide to take part you will be given this information sheet to keep and asked to sign a consent form. You are free to withdraw at any time without giving a reason. You also have the right to withdraw retrospectively any consent given, and to require that any data gathered on you be destroyed.

The experiment will last up to 1.5 hours and you will receive a reward of £8 upon completion. You will first be asked to provide demographic information in an introductory questionnaire. In order to identify news topics of your interest, we provide you a list of the most important news reports which have been reported on the BBC News Website between November 2008 and April 2009. Your task will be to select those news articles which interest you. Your second task will be to categorise the selected articles into related groups. If possible, please choose rather broad descriptions for each category without using too general descriptions. For instance, if you selected an article about the US Presidential Election Campaign, you might want to use the label “*US Politics*” instead of “*Politics*” or “*US Presidential Election Campaign*” to categorise it. After this task, you will be asked to fill in a Post-Task Questionnaire to gain an insight into your impressions of this task.

Based on the defined categories of the previous task, we will then provide you a list of news video stories which have been broadcast on public television during the same time period covered in your first task. Your next task will be to mark the relevance of each displayed news story to the given category on a scale from 0 (not relevant) to 5 (highly relevant). At the end of this task, you will be asked to fill in an Exit Questionnaire. Please ask questions if you need to and remember that you can opt out at any time during the experiment.

All information collected about you during the course of this study will be kept strictly confidential. You will be identified by an ID number and all information about you will have your name and contact details removed so that you cannot be recognised from it. Data will be stored for analysis, and then destroyed. The results of this study may be used for some PhD research. You will not be identified in any report or publication that arises from this work.

This study is being funded by the European Salero project at the Department of Computer Science, University of Glasgow. This project has been reviewed by the Faculty of Information and Mathematical Sciences Ethics Committee.

For further information about this study please contact:

Frank Hopfgartner
Department of Computing Science, University of Glasgow
18 Lilybank Gardens
Glasgow, G12 8QQ
Email: hopfgarf@dcs.gla.ac.uk
Tel.: 0141 330 1641

TECHNICAL INFORMATION SHEET



Project: Creating personalised ground truth data

Researcher: Frank Hopfgartner

**UNIVERSITY
of
GLASGOW**

Your task will be to select from a list of news articles those stories that interest you. Moreover, you will be asked to categorise these stories into broad groups. Based on these categories, we will then provide you a list of video news stories which have been broadcast on public television. Your task will be to rate the relevance of these stories to the given category. In this information sheet, we provide you some technical information you might find useful to understand your tasks.

BBC News Article Judgement:

Your first task is to select news events, represented by according online news articles, you are interested in. In this work, we focus on news reports covering events happening between November 2008 and April 2009. We downloaded, using the Google Search API, the top ten relevant articles of each working day from the BBC News website within this time span.

We split these stories into blocks of 14 days. For each block, we provide you a web page containing the title and a short abstract of each story. Clicking on the title opens the original article in another browser window. Please inspect each article carefully and mark those articles as relevant by ticking the box next to each story. Once you selected all according articles, click the “Submit” button on the bottom of the page.

Make sure that you entered the User ID which we will assign to you on the top of the page before submitting your judgements.

The next page you will see lists all stories which you selected on the previous page. Again, you can inspect the articles further by clicking on the according titles. Your task here will be to categorise each story belonging to a broader news topic. If possible, please choose rather broad descriptions for each category without using too general descriptions. For instance, if you selected an article about the US Presidential Election Campaign, you might want to use the label “*US Politics*” instead of “*Politics*” or “*US Presidential Election Campaign*” to categorise it. A click on “Submit” will lead you to the next block of articles. Please repeat the illustrated tasks until you judged all remaining news stories.

Video News Story Judgement:

Between November 2008 and April 2009, we captured the daily broadcast of the BBC One O’Clock News and the ITV Evening News. These bulletins have been automatically split into news video stories. The feedback you provided in the previous task results in a list of news aspects you are interested in. Examples could be “US Politics”, “Rugby 6 Nations Cup” or “Unemployment in UK”.

Your next task is to identify video stories for each of these aspects using a web based interface. In this interface, results (stories) for each aspect are represented by example key frames and by a textual transcript. You can browse through the key frames by clicking on them. Moreover, you can play each video by clicking on the small playback icon.

Please judge for each displayed story if it is related to the given aspect which is labelled on top of the search results. You can judge the relevance from 0 (not relevant) to highly relevant (5) using the individual scale bars. **Note that results for each aspect are split into various result pages. You can navigate to the next results page by clicking on the according numbers on the top or the bottom of the result list.** Once you finished evaluating all presented stories, a click on “Next aspect’s results” will display stories of the next aspect you defined in the previous task. Please repeat the judgment for every presented story for each aspect.

CONSENT FORM

Project: Creating personalised ground truth data

Researcher: Frank Hopfgartner



**UNIVERSITY
of
GLASGOW**

Please tick box

1. I confirm I have read and understand the information sheet for the above study and have had the opportunity to ask questions. ☐
2. I understand that my permission is voluntary and that I am free to withdraw at any time, without giving any reason, without my legal rights being affected. ☐
3. I agree to take part in the above study. ☐
4. I would like to receive a summary sheet of the experimental findings ☐

If you wish a summary, please leave an email address _____

Name of Participant Date Signature

Researcher Date Signature



**UNIVERSITY
of
GLASGOW**

Entry Questionnaire

* Required

User ID: *

Please ask the experimenter to fill this in.

Please indicate your gender. *

- ☐ male
- ☐ female

Please provide your age. *

Please provide your current occupation. *

What is your educational level? *

- ☐ No degree
- ☐ Undergraduate
- ☐ Postgraduate
- ☐ Other:

What is your field of work or study? *

Please indicate which media types you use to receive latest news. *

Mark AS MANY as possible.

- ☐ Television
- ☐ Radio (e.g. in the car)
- ☐ Newsfeeds/Newsletters (e.g. Google Alerts)
- ☐ News Media Websites (e.g. BBC)
- ☐ Word-of-mouth
- ☐ Other:

Please describe your news consumption habits. *

(e.g. watching TV after dinner, or checking news websites every hour,...)

B.4. Entry Questionnaire

Please select news topics of your interest. *

- ☐ Business & Finance
- ☐ Entertainment & Culture
- ☐ Health, Medical & Pharma
- ☐ Politics
- ☐ Sports
- ☐ Technology & Internet
- ☐ Other:

Please name sub topics for each of the above selected topics of interest. *

e.g. Football and/or Rugby as sub categories of Sports

B.5. Post-Task Questionnaire

Post-Task Questionnaire

Thank you for your assessment. In order to review your judgments, we now ask you to answer some questions about it. Take into account that we are interested in knowing your opinion: answer questions freely, and consider there are no right or wrong answers.

* Required

User ID *

If you are unsure, please ask the experimenter to fill this in.

Judging the relevance of articles was generally *

1 2 3 4 5

Simple ☐ ☐ ☐ ☐ ☐ Complex

If you found this task complex, why do you think the task was complex?

Before starting the task, I had a general idea of which news events happened in the given time period *

1 2 3 4 5

Disagree ☐ ☐ ☐ ☐ ☐ Agree

Before starting the task, I knew which kind of stories I was interested in. *

1 2 3 4 5

Agree ☐ ☐ ☐ ☐ ☐ Disagree

During the task, I discovered interesting news events which I was not aware of before *

1 2 3 4 5

Disagree ☐ ☐ ☐ ☐ ☐ Agree

I marked various news events as interesting even though I was not interested in them at the given time period. *

1 2 3 4 5

Agree ☐ ☐ ☐ ☐ ☐ Disagree

If you marked these events as interesting, please state why.

Additional comments

(Optional)

B.6. Exit Questionnaire

Exit Questionnaire

Thank you for your assessment. Please fill in this final exit questionnaire so we can evaluate your judgments.

* Required

User ID: *

If you are unsure, please ask the experimenter to fill this in.

Overall, the displayed news stories were related to the according aspect *

1 2 3 4 5

Agree ☐ ☐ ☐ ☐ ☐ Disagree

The displayed news stories covered most facets of the according aspect *

1 2 3 4 5

Disagree ☐ ☐ ☐ ☐ ☐ Agree

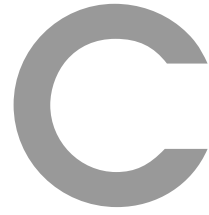
The story segmentation was appropriate *

1 2 3 4 5

Agree ☐ ☐ ☐ ☐ ☐ Disagree

Additional comments

(optional)



Semantic User Modelling for Personalised News Video Access: Experimental Documents

This appendix presents the experimental documents described in Section 7.3. These include:

- C.1: Information Sheet
- C.2: Simulated Work Task Situation
- C.3: Consent Form
- C.4: Entry Questionnaire
- C.5: Interim Questionnaire
- C.6: Exit Questionnaire

Note that the participants were asked to fill in questionnaires (C.4 – C.6) online.

Information Sheet

Project: Semantic User Modelling for Personalised News Video Access

Researcher: Frank Hopfgartner



**UNIVERSITY
of
GLASGOW**

You are invited to take part in a research study. Before you decide to do so, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully. Ask me if anything is not clear or if you would like more information. It is up to you to decide whether or not to take part. If you decide to take part you will be given this information sheet to keep and asked to sign a consent form. You are free to withdraw at any time without giving a reason. You also have the right to withdraw retrospectively any consent given, and to require that any data gathered on you be destroyed.

The results of this study will be used for research publications. The aim of this evaluation is to compare the performance of two different video recommender systems. You will be randomly assigned to use one of the recommender systems for the duration of the evaluation. The aim of this research is to study long term user profiling and personalized news video recommendation. We are performing a multi-session interactive information retrieval evaluation, which requires user interactions under realistic conditions. You will therefore be asked to include the recommender system into your daily news gathering routine for up to two weeks. Besides using the system to satisfy your personal information need, you will be asked to perform a simulated work simulation task over each session. We require you to use the system for at most ten minutes each day until the end of the study.

The maximum total duration of this experiment is 2 hours and 25 minutes across two weeks. You will be paid £15 upon completion of the experiment and a reduced rate if you do not complete the experiment. All information collected about you during the course of this study will be kept confidential. You will be identified by an ID number and all information about you will have your name and contact details removed so that you cannot be recognized from it. Data (i.e. key strokes, questionnaire feedback and mouse clicks) will be stored for analysis, and then destroyed.

You will first be asked to complete an Entry Questionnaire to obtain group demographics. Then, you will be given a short tutorial to familiarize yourself with the video recommender system. The search sessions will be performed independently, at a computer you can easily access in your home or a public place. With your permission, the experimenter will email you on the morning of each day to remind you about the evaluation. You will be asked to complete a Post-Session Questionnaire after the 2nd, 4th, 6th, 8th and 10th search session to gain an insight into your impressions of the system. Finally, you will be asked to complete an Exit Questionnaire.

The evaluation has been reviewed by the Faculty of Information and Mathematical Sciences Ethics Committee.

For further information about this study please contact:

Frank Hopfgartner
Department of Computing Science, University of Glasgow
AW220, Sir Alwyn Williams Building
Glasgow, G12 8RZ
Email: hopfgarf@dcs.gla.ac.uk
Tel.: 0141 330 1641



Simulated Work Task Situation

Project: Semantic User Modelling for Personalised News Video Access

Researcher: Frank Hopfgartner



**UNIVERSITY
of
GLASGOW**

“You just started a new job in a remote town away from home. Since you do not know anyone in town yet, you are keen to socialise with your colleagues. They seem to be big sports supporters and are always up for some sports related small talk. Sport hence opens great opportunities to start conversations with them. Luckily, there are various major sports events and tournaments this month which they all show interest in, e.g., the Winter Olympics in Vancouver, the Rugby Six Nations Cup and European Football Tournaments. Every day, you eagerly follow every news report on the BBC to be up to date and to join their conversations.”

Indicative Request:

You should use the recommender system to follow sports related news. This might include major international events such as the Winter Olympics, European football competitions or the Rugby Six Nations cup. Reports might be summaries of the competition day, feature reports about Team GB, or summaries of football/rugby matches.

Keep in mind that you should follow the news well enough to be able to chat and socialise with your new colleagues.

CONSENT FORM

Project: Semantic User Modelling for Personalised News Video Access

Researcher: Frank Hopfgartner



Please tick box

1. I confirm I have read and understand the information sheet for the above study and have had the opportunity to ask questions.

☐

2. I understand that my permission is voluntary and that I am free to withdraw at any time, without giving any reason, without my legal rights being affected.

☐

3. I agree to take part in the above study.

☐

4. I authorise the researcher to send me daily emails to remind me to continue this study.

☐

5. I would like to receive a summary sheet of the experimental findings

☐

My email address is

Name of Participant

Date

Signature

Researcher

Date

Signature

Entry Questionnaire

This questionnaire will provide us with background information that will help us analyse the answers you give in later stages of this experiment. You are not obliged to answer a question if you feel it is too personal. The information is kept completely confidential and no information is stored on computer media that could identify you as a person.

User ID

Please ask the experimenter to fill this in.

Please indicate your GENDER

- ☐ male
☐ female

Please provide your AGE

- ☐ 18-25
☐ 26-30
☐ 31-39
☐ >40

Please provide your current OCCUPATION

What is your educational level

- ☐ No degree
☐ Undergraduate student
☐ Graduate student (PhD, MSc)
☐ Faculty/Research Staff
☐ Other:

What is your field of work or study?

How would you describe your proficiency with ENGLISH?

1 2 3 4 5

Beginner ☐ ☐ ☐ ☐ ☐ Native Speaker

Please indicate which online search services you use to retrieve MULTIMEDIA content.

Mark AS MANY as suitable.

- ☐ Google (<http://www.google.com>)
☐ Yahoo! (<http://www.yahoo.com>)
☐ AltaVista (<http://www.altavista.com>)
☐ YouTube (<http://www.youtube.com>)
☐ Flickr (<http://www.flickr.com>)
☐ Microsoft (<http://www.live.com>)
☐ Other:

C.4. Entry Questionnaire

Please indicate which media type you use to receive LATEST NEWS

Mark AS MANY as suitable.

- ☐ Television
- ☐ Radio (e.g. in the car)
- ☐ Newsfeeds/Newsletters (e.g. Google Alerts)
- ☐ News Media Webpages (e.g. BBC iPlayer)
- ☐ Word-of-mouth
- ☐ Other:

Please select news topics of your INTEREST

Mark AS MANY as suitable.

- ☐ Business & Finance
- ☐ Entertainment & Culture
- ☐ Health, Medical & Pharma
- ☐ Politics
- ☐ Sports
- ☐ Technology & Internet
- ☐ Other:

Describe your usual news consumption habit

(e.g. immediate consumption, late night consumption,...)



Interim Questionnaire

To evaluate the system you have used, we now ask you to answer some questions about it. Take into account that we are interested in knowing your opinion: answer questions freely, and consider there are no right or wrong answers. Please remember that we are evaluating the system you have just used and not you.

* Required

UserID: *

Please indicate what you used the system for *

- ☐ Finding videos of older news
- ☐ Identifying latest news
- ☐ Identifying new stories you haven't heard of before
- ☐ Other:

Please indicate the news categories you were interested in in the last few days. *

- ☐ Business & Finance
- ☐ Entertainment & Culture
- ☐ Health, Medical & Pharma
- ☐ Politics
- ☐ Sports
- ☐ Technology & Internet
- ☐ Other:

These categories were successfully identified and displayed on the left hand side of the interface *

1 2 3 4 5

Agree ☐ ☐ ☐ ☐ ☐ Disagree

Using the system, it was easy to retrieve stories belonging to these categories *

1 2 3 4 5

Agree ☐ ☐ ☐ ☐ ☐ Disagree

The displayed sub categories represent my diverse interests in various topics. *

1 2 3 4 5

Disagree ☐ ☐ ☐ ☐ ☐ Agree

The displayed results for each sub category were related to each other *

1 2 3 4 5

Agree ☐ ☐ ☐ ☐ ☐ Disagree

The displayed results for each category matched with the category description. *

1 2 3 4 5

Disagree ☐ ☐ ☐ ☐ ☐ Agree

C.5. Interim Questionnaire

The displayed results for each category contained relevant stories I didn't retrieve otherwise. *

12345

Agree ☐ ☐ ☐ ☐ ☐ Disagree

The system was effective in automatically identifying my interests *

12345

Agree ☐ ☐ ☐ ☐ ☐ Disagree

The interface structure helped me to explore the news collection. *

12345

Agree ☐ ☐ ☐ ☐ ☐ Disagree

The system helped me to explore various topics of interest *

12345

Agree ☐ ☐ ☐ ☐ ☐ Disagree

Additional comments
(Optional)

Exit Questionnaire

Please consider the entire search experience that you just had when you respond to the following question.

* Required

UserID: *

I mainly used the system to perform the pre-defined search topic *

1 2 3 4 5

Agree ☐ ☐ ☐ ☐ ☐ Disagree

The system was a useful addition to my daily news gathering process *

1 2 3 4 5

Agree ☐ ☐ ☐ ☐ ☐ Disagree

I would use commercialised systems like that for my daily news gathering *

1 2 3 4 5

Disagree ☐ ☐ ☐ ☐ ☐ Agree

What did you like about the system? *

What did you DISLIKE about the system? *

Additional comments